IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024PAP0003

Publicized:2024/06/26

A PUBLICATION OF THE INFORMATION AND SYSTEMS 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

Design and implementation of opto-electrical hybrid floating-point multipliers

Takumi INABA[†], ***Nonmember***, **Takatsugu ONO**[††], **Koji INOUE**[††], ***and*** **Satoshi KAWAKAMI**[††], ***Members***

SUMMARY

The performance improvement by CMOS circuit technology is reaching its limits. Many researchers have been studying computing technologies that use emerging devices to challenge such critical issues. Nanophotonic technology is a promising candidate for tackling the issue due to its ultra-low latency, high bandwidth, and low power characteristics. Although previous research develops hardware accelerators by exploiting nanophotonic circuits for AI inference applications, there has never been considered for the acceleration of training that requires complex Floating-Point (FP) operations. In particular, the design balance between optical and electrical circuits has a critical impact on the latency, energy, and accuracy of the arithmetic system, and thus requires careful consideration of the optimal design. In this study, we design three types of Opto-Electrical Floating-point Multipliers (OEFMs): accuracy-oriented (Ao-OEFM), latency-oriented (Lo-OEFM), and energy-oriented (Eo-OEFM). Based on our evaluation, we confirm that Ao-OEFM has high noise resistance, and Lo-OEFM and Eo-OEFM still have sufficient calculation accuracy. Compared to conventional electrical circuits, Lo-OEFM achieves an 87% reduction in latency, and Eo-OEFM reduces energy consumption by 42%.

key words: *Opto-Electrical circuit, analog computing, floating-point multiplier, silicon photonics*

1. Introduction

The end of Dennard scaling has led to the development of dedicated hardware accelerators for highly efficient execution. However, from a long-term perspective, there are limits to improving the performance achieved by CMOS circuits because we cannot expect sustainable transistor shrinking, i.e., the end of Moore's Law. Many researchers have been studying computing technologies that take advantage of emerging devices to address such critical issues. Nanophotonic technology is a promising candidate due to its ultra-low latency, high bandwidth, and low power natures.

Although nanophotonics computing has demonstrated outstanding potential for AI inference applications [1, 2], there has never been considered for the acceleration of training that requires complex Floating-Point (FP) operations with exponent and mantissa handling, digit alignment, rounding functions, etc. This situation makes implementing an all-optical design extremely difficult. A promising direction is introducing an Opto-Electrical hybrid style, i.e., exploiting the ultra-low-latency optical integer units with complex electrical data management to form an FP unit. In this case, the main challenges are as follows. First, the number of optical and electrical boundaries should be minimized. This is because the optical and electrical elements work in the analog and digital domains, respectively, requiring not only optical-electrical but also analog-digital conversions. Second, although applying optical circuits aggressively reduces the number of boundaries, on the other hand, it worsens the computing accuracy due to the noise-sensitive analog operations. Unfortunately, as far as we know, the design of Opto-electric hybrid FP arithmetic units has never been discussed, and the impact of the hybridization strategy on energy efficiency and calculation accuracy is unclear.

In this paper, we target FP multipliers, which is a key component to achieving optically-accelerated AI training [†]. The contributions of this work are as follows.

- We identify the FP multiplier's latency and energy consumption bottlenecks. This analysis helps determine which parts of the FP multiplier should be optically implemented.
- Optical components are proposed, such as a round unit required to explore and form opto-electrical hybrid FP multipliers.
- Three types of Electrical hybrid Floating-point Multipliers (OEFMs) using the introduced optical components, accuracy-oriented (Ao-OEFM), latency-oriented (Lo-OEFM), and energy-oriented (Eo-OEFM), are designed.
- Based on our evaluation, we confirm that Ao-OEFM has high noise resistance, and Lo-OEFM and Eo-OEFM still have sufficient calculation accuracy.
- Compared to conventional electrical circuits, Lo-OEFM achieves an 87% reduction in latency, and

[†]The author is with the Graduate School of Information Science and Electrical Engineering, Kyushu University

[††]The author is with the Faculty of Information Science and Electrical Engineering, Kyushu University

[†]The initial design (Ao-OEFM in Fig. 2) has reported in [3], i.e., only the integer multiplier unit is implemented in an optical circuit. In addition to the initial design, this paper designs Lo-OEFM and Eo-OEFM shown in Fig. 2 as other design alternatives in order to explore the Opto-Electrical hybrid FP design.

Eo-OEFM reduces energy consumption by 42%.

The paper is structured as follows: Section 2 presents the current status of optical arithmetic units. In Section 3, we detail the three proposed OEFM designs, including the optical devices' integer multiplier and adder. Section 4 outlines the evaluation framework, while Section 5 presents experimental results and discusses the advantages of the optical arithmetic unit. Finally, Section 6 concludes the paper.

2. Basics of floating-point arithmetic and optical computing

2.1 Floating-point arithmetic overview

Machine learning has been actively applied to numerous fields thanks to the continuous development of computers. In machine learning, the primary process during training is the FP sum-of-products operation. Since the low latency and energy consumption of FP arithmetic directly impact the efficiency of computer systems for machine learning, extensive research has been conducted. Deep neural networks have been successfully trained using 8-bit FP numbers while maintaining accuracy [4]. A low-cost hardware implementation using Bfloat16-square integration has been reported [5].

Bfloat16, an FP representation format for machine learning, was standardized by Google Inc. FP notation comprises three parts: Sign, Exponent, and Fraction. In Bfloat16, the Sign part is 1-bit, the Exponent part is 8-bits, and the Fraction part is 7-bits. Bfloat16 incorporates a bias value of 127, which is added to the Exponent part. Various ongoing studies regarding Bfloat16 encompass hardware performance evaluation [6] and the development of binary analysis tools [7]. However, to the best of the author's knowledge, there is limited research on FP arithmetic units that support Bfloat16. Among the FP arithmetic units—adders, multipliers, and dividers—this study primarily focuses on a floating-point multiplier that supports Bfloat16 as its first step. A low-latency and energy-efficient multiplier supporting Bfloat16 is anticipated to contribute to energy-efficient machine learning with reduced latency.

2.2 Optical computing: Opportunities and challenges

With the end of Dennard scaling, computing with novel devices is attracting significant attention to achieve higher performance and lower energy consumption for arithmetic units. Multicore scaling has been found to be power-limited, irrespective of chip configuration or topology [8]. Research using novel devices is diverse, including the use of plants [9], a quantum microarchitecture [10], and a superconducting single-flux quantum device [11]. However, these innovative computing systems have extremely severe environmental constraints (e.g., cryogenic temperatures). Optical devices are often used in communication technology and are not limited by environmental conditions. In other words, computing with optical devices is one of the most promising technologies that could become commonplace.

Some research has been conducted to date toward the development of light-based digital and analog arithmetic. In digital units, all-optical logic gates using semiconductor optical amplifiers (SOAs) [12] and logic gates utilizing the light beam interference effect [13] have been created. Combining optical logic gates creates half-adder [14] and full-adder [15]. In analog units, the potential advantages of a photonic accelerator (PAXEL) and the scope for future work toward practical implementation have been reported [16]. Highly efficient differential and integral calculations using the spatial Fourier transform concept have also been highlighted [17]. Realizing FP arithmetic using optical devices may achieve low latency and energy consumption compared with CMOS devices. However, applying the current optical arithmetic unit to an FP arithmetic unit is difficult. Digital optical circuits have not reported such complex arithmetic units as FP arithmetic. In analog optical circuits, the representation space of FPs far exceeds what values analog arithmetic units can achieve. Therefore, we propose OEFM using analog-based optical and digital-based electrical arithmetic units. This method can potentially combine the benefits of low latency and energy consumption from optical devices with the high precision operations from electrical devices. In proposing OEFM, the balance between optical-analog and electrical-digital arithmetic must be considered. Optical arithmetic units introduce a trade-off between accuracy and latency/energy consumption. Therefore, multiple design patterns need to be evaluated. To our knowledge, our work is the first study of an FP multiplier utilizing an optical device.

2.3 Performance/energy impact for FP multiplier

When designing an OEFM, determining the allocation of tasks between optical-analog and electrical-digital arithmetic is crucial. Optical analog arithmetic units offer low latency and energy consumption benefits, but there is a trade-off with reduced arithmetic accuracy. Additionally, optical analog arithmetic involves converters (analog-to-digital converter (ADC), digital-to-analog converter (DAC), optic-to-electro converter (OEC), and electro-to-optic converter (EOC)), introducing latency and energy consumption that may offset the advantages. Therefore, in OEFM design, it is vital to judiciously incorporate optical-analog operations where they can be effectively utilized rather than indiscriminately increasing their use.

To achieve low latency and energy consumption of FP multipliers, we estimate the number of logic gates and gate depths for each component in electrical-digital

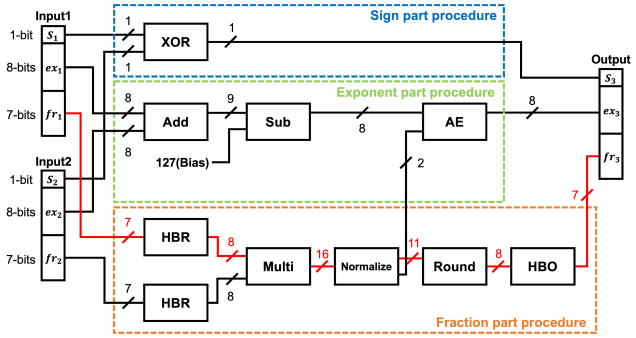


Fig. 1 FP multiplier components and configuration

circuits on a model basis. On the basis of the analysis results, the components that should be replaced with optical analog circuits are discussed. Figure 1 shows a circuit diagram of the FP multiplier, consisting of the Sign, Exponent, and Fraction parts. The Sign part is calculated by XORing the Sign part of input1 and input2. The Exponent part includes Add, Sub, and Adjust Exponent (AE). The Exponent part of input1 and input2 is added, and then the bias is subtracted from their result in Sub. If normalization is required, the Exponent part is corrected by adding 1 to the Exponent part in AE and dividing the Fraction part by 2 in Normalized. The Fraction part includes Hidden Bit Restore (HBR), Multi, Normalized, Round, and Hidden Bit Omit (HBO). The Fraction part omits the hidden bit, which is restored in HBR. The Fraction part of input1 and input2 is multiplied in Multi. In Normalized, the output of Multi is normalized. In Round, the Fraction part is rounded to 8-bits. This research assumes "round to nearest - even" as the rounding approach. Finally, the hidden bit in HBO is omitted. Regarding accuracy, Sign and Exponent calculations must be more accurate than Fraction part calculations. Incorrect Sign or Exponent calculations may lead to a shift from positive to negative or result in a doubled value, causing a complete deviation from the correct outcome.

This research utilizes a model [18] [19] to estimate the number of gates and gate depths for each component. The number of gates is directly related to energy consumption, while gate depth is proportional to latency. Estimation was performed for the following electric floating-point multipliers. In this research, priority is given to low-latency circuits, and if the latency of the entire floating-point multiplier does not change, low-energy circuits are selected. Add and Sub are designed using Ripple Carry Adder (RCA), and Multi employs an array multiplier with Carry Save Adder (CSA) [18] [19]. Additionally, Carry Look-ahead Adder (CLA) [18] is used as an adder in Round and AE.

Table 1 shows the number of gates and depth for each component. By focusing on gate depth, or latency, we can see that Multi, Add, Sub, Round, and AE have large delays. Floating-point multipliers have

Table 1 Breakdown of gate count and depths for FP multiplier

Component name	Gate count (Breakdown)	Gate depths
XOR	3 (0.23%)	2
Add	72 (5.54%)	16
Sub	72 (5.54%)	16
Multi	708 (54.5%)	35
Normalized	17 (1.31%)	6
AE	212 (16.3%)	10
Round	215 (16.6%)	13
HBR	0 (0%)	0
HBO	0 (0%)	0
Total	1299 (100%)	–

three paths from input to output. The path that takes the longest is the critical path, which determines the overall delay. The thick line in Figure 1 is the critical path. Next, by focusing on the number of gates, or the energy consumption, we can see that the energy consumption of Multi, Round, AE, Add, and Sub is large.

From the analysis results, XOR, HBR, and HBO are not considered for optical implementation because of their negligible impact on latency and energy (rather, latency and energy consumption may increase due to converter overhead). Normalized is challenging to implement in analog because it involves digital concepts ("round to nearest - even"). We therefore considered whether it is better to do optical analog or electrical digital for each of Add, Sub, AE, Multi, and Round. We design OEFM with three focuses: accuracy, latency, and energy consumption.

3. Opto-electrical FP multipliers design

3.1 Overview

In OEFM, optical analog arithmetic is computed using several optical devices. The laser is an almost ideal monochromatic light source. The laser output is represented by Equation (1).

$$E = Ae^{j(\omega t+\theta)} \tag{1}$$

Where A is the electrical field amplitude of light, ω is the angular frequency, t is the time, and θ is the initial phase. In this paper, A represents an information carrier. We design three optical arithmetic units: Optical-Multi, Optical-AddSubAE, and Optical-Round. Optical-Multi is a new optical integer multiplier. Optical-Multi consists of a laser, phase shifter, X-coupler, photodiode, DAC, and ADC. Optical-AddSubAE (Figure 3) and Optical-Round are simple adders based on the superposition principle. Optical-AddSubAE and Optical-Round consist of a laser, phase shifter, photodiode, DAC, and ADC.

3.2 OEFM design choises

Figure 2 shows the three types of OEFMs proposed in this research: accuracy-oriented (Ao-OEFM)†, latency-

†Ao-OEFM is the design presented in [3], and Lo-OEFM

	Ao-OEFM①	Lo-OEFM②	Eo-OEFM③
XOR	E	E	E
Add	E	O	O
Sub	E	O	O
Multi	O	O	O
Normalized	E	E	E
AE	E	O	O
Round	E	O	E
HBR	E	E	E
HBO	E	E	E

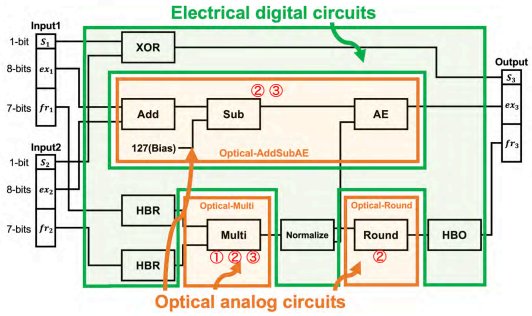


Fig. 2 OEFM design choises

oriented (Lo-OEFM), energy-oriented (Eo-OEFM). In Ao-OEFM, Multi is an optical analog circuit; the other components are electrical circuits. In Lo-OEFM, Multi, Add, Sub, Round, and AE are optical circuits; the other components are electrical circuits. In Eo-OEFM, Multi, Add, Sub, and AE are optical circuits; the other components are electrical circuits. The optical analog circuit includes a DAC, EOC, OEC, and ADC.

3.3 Optical arithmetic units for OEFMs

3.3.1 Optical devices

The phase shifter generates a phase shift. E_{in} and E_{out} are the input and output lights respectively (where E corresponds to the equation shown in equation (1)). This research uses a $-\pi/2$ and π phase shifter; the outputs are shown in Equations (2) (3), respectively

$$E_{out} = -jE_{in} \tag{2}$$

$$E_{out} = -E_{in} \tag{3}$$

The X coupler is a 2-input, 2-output device that combines and splits optical signals. Half of the light to the input port goes to the opposite output port, and the remainder goes straight to the other port. As it travels to the opposite output port, the phase is shifted by $+\pi/2$. The X coupler's transmission matrix to the cross is shown in Equation (4). E_{in1} is the upper input, E_{in2} is the lower input after PS, E_{out1} is the upper output, and E_{out2} is the lower output.

and Eo-OEFM are newly designed in this paper.

$$\begin{pmatrix} E_{out1} \\ E_{out2} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix} \begin{pmatrix} E_{in1} \\ E_{in2} \end{pmatrix} \tag{4}$$

The photodiode converts an optical signal to an electrical current, with the output current value being the square of the input light's electrical field amplitude.

3.3.2 Optical-Multi

The implementation of Optical-Multi is proposed in [3], and we employ it in this paper. Let the two inputs to the Optical-Multi be input data 1 and input data 2, which are the electrical field amplitude A_1 of E_{in1} (the upper laser) and the field amplitude A_2 of E_{in2} (the lower laser), respectively. When the frequency and initial phase are the same, the two laser lights are defined by Equation (5).

$$\begin{pmatrix} E_{in1} \\ E_{in2} \end{pmatrix} = \begin{pmatrix} A_1 e^{j(\omega t+\theta)} \\ A_2 e^{j(\omega t+\theta)} \end{pmatrix} \tag{5}$$

Light passing through the PS and X couplers is received by two photodiodes connected so that the current flows in the opposite direction, respectively. When the current values converted by the photodiode are I_1 and I_2, the current value I_{out}, when connected in the opposite direction, is expressed by Equation (6).

$$\begin{aligned} I_{out} &= I_1 - I_2 \\ &= \frac{1}{2}(A_1^2 + A_2^2 + 2A_1A_2) - \frac{1}{2}(A_1^2 + A_2^2 - 2A_1A_2) \\ &= 2A_1A_2 \end{aligned} \tag{6}$$

The output of ADC can be A_1A_2 by setting the threshold interval in the ADC. Therefore, a circuit whose output current value is $2A_1A_2$ can function as a multiplier. For more detailed principles and implementation, please refer to [3].

3.3.3 Optical-AddSubAE and Optical-Round

In Figure 3, the four inputs of Optical-AddSubAE are represented by Equation (7).

$$\begin{pmatrix} E_{in1} \\ E_{in2} \\ E_{bias} \\ E_{carry} \end{pmatrix} = \begin{pmatrix} A_1 e^{j(\omega t+\theta)} \\ A_2 e^{j(\omega t+\theta)} \\ 127 e^{j(\omega t+\theta)} \\ A_{carry} e^{j(\omega t+\theta)} \end{pmatrix} \tag{7}$$

After the bias passes through the PS with a π shift, the three lights interfere with the waveguide. Let E_{out} be the result of this interference, and E_{out} is expressed by Equation (8).

$$\begin{aligned} E_{out} &= E_1 + E_2 + E_{bias} + E_{carry} \\ &= (A_1 + A_2 - 127 + A_{carry})\, e^{j(\omega t+\theta)} \end{aligned} \tag{8}$$

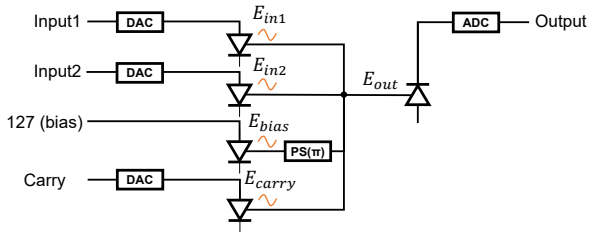


Fig. 3 Optical-AddSubAE

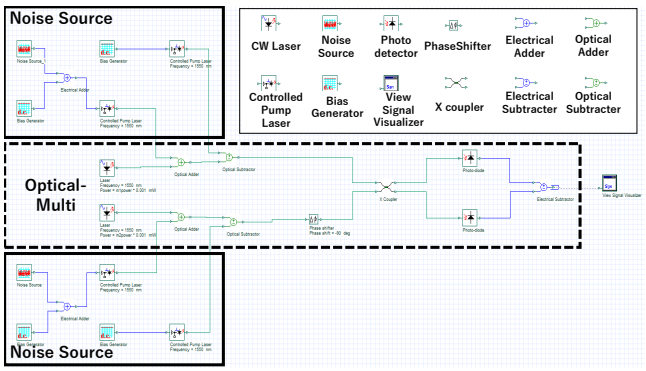


Fig. 4 Circuit schematic in OptiSystem

Thus, in optical-analog circuits, addition and subtraction can function with simple interference on the basis of the superposition principle. Adders in Optical-Round can function with simple interference as well as Optical-AddSubAE.

4. Experimental set up

4.1 Purpose of experiment

The experiment aims to compare the arithmetic accuracy, latency, and energy consumption of the three OEFMs and the Electrical-FM (baseline), which are all electrical components. Since optical arithmetic units involve a trade-off between accuracy and latency/energy consumption, we take the following two steps for the evaluation. As the first step, we confirm the noise impact on the OEFM accuracy, assuming a wide noise range (10^{-15} to 10^{-3} [mW]) for comprehensive sensitivity analysis. Although there is a concern that noise generated in optical devices could reduce arithmetic accuracy, our evaluation results demonstrate that this drawback is negligible under realistic design parameter settings (Section 5.1). Optical arithmetic units have a trade-off between error rate and energy consumption depending on the laser light intensity, i.e., higher intensity improves the arithmetic accuracy by consuming more energy. It has been observed that by providing enough level of laser light intensity, we can achieve a sufficiently large signal-to-noise ratio (SNR) that makes the error rate negligible even if we assume a realistic noise level. Based on such observation, as the second step, we perform iso-accuracy latency/energy comparison for OEFMs with the Electrical-FM baseline. We assume the design parameters for the OEFMs used in the accuracy analysis, i.e., ensuring enough laser light intensity for error-free computation, and have found significant latency/energy advantages of the OEFM designs over the full-electric baseline (Section 5.2).

This experiment is valuable for considering the balance between optical analog and electrical digital arithmetic. The accuracy survey clarifies the noise effects on the optical arithmetic unit and OEFMs output. One of the causes of arithmetic errors is the noise that occurs in optical arithmetic unit components. The latency/energy consumption experiment calculates the latency/energy consumption of three OEFMs and an Electrical-FM on the model. The introduction of optical arithmetic units incurs the overhead of converters, making it only sometimes possible to achieve low latency and energy consumption.

Additionally, since AI computation is a potential target for optical acceleration, throughput is also an important consideration. Electrical circuits such as Electrical-FM can generally be executed in parallel by pipelined circuit slicing to increase throughput. We discuss the potential benefits of OEFM by comparing the ideally pipelined Electrical-FM and OEFM in terms of throughput and energy. To make a pessimistic evaluation, we ignore the overheads of pipelining except for the pipeline register and consider a situation where no pipeline stall occurs, i.e., we ignore the overheads such as the pipeline controller unit and wiring. We assume that the operating frequency is proportional to the number of pipeline stages and that the only increase in energy consumption is the addition of pipeline registers.

4.2 Experimental environment for evaluating the arithmetic accuracy

4.2.1 Experiments with the Optical arithmetic units

We implement the optical arithmetic units on OptiSystem [20], version 21.0.0, a software simulator for designing and verifying optical systems. Figure 4 shows Optical-Multi designed within OptiSystem. In the experiment, a virtual device called a noise source collectively generates noise while other devices remain noise-free. The NoiseSource reproduces the accumulation of noise generated by each device and is placed immediately after the laser. The circuit on OptiSystem includes two lasers, a PS, an X coupler, two photodiodes, and two Noise Sources. The DAC and ADC are not

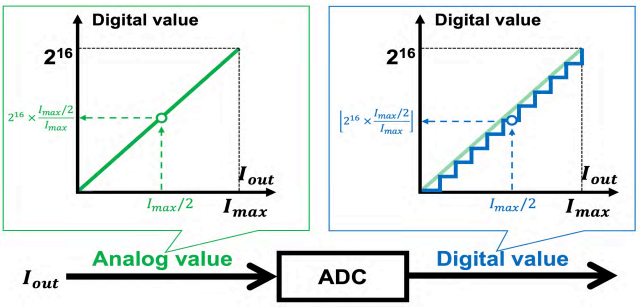


Fig. 5 Decoding of analog and digital values

included in the circuit on OptiSystem.

The data flow is from left to right. The input to the Optical-Multi is reflected in the field amplitude of the laser light. Next, the Noise Sources add noise to the laser light. The output of the photodiode is an analog current value and is detected by the view signal visualizer. As described in Section 3.3.2, the angular frequency ω and initial phase θ of the two laser beams match.

The input data sample (Input1 and Input2 pairs) is a 16-bit random sequence. This ensures that positive and negative numbers, as well as large and small numbers in the Exponent part, are used as samples evenly. The sample data is random, but the same sample set is used for each experiment with different noise variances.

An Analog value is defined to examine the effect of noise on the output of the OEC. The Analog value is converted from the analog current value and could be evaluated for accuracy. I_{out} is the analog current value of the output of the OEC. I_{out} is linearly transformed so that 2^{16} corresponds to the maximum analog current value I_{max}. This means that the analog current value at the output of the OEC is converted to a 16-bit value, which is the output of the ADC, up to the decimal point, to indicate how much the value is. Figure 5 shows an outline of the linear transformation.

When both DAC outputs are their maximum of 255, the theoretical I_{max} is 130.05 [mA]. $255 \times 255 = 65025$, so given a proportion such that 130.05 [mA] and 65025 correspond, $130.05 \times 500 = 65025$, so the value in I_{out} [mA] $\times$ 500 corresponds to the output of the ADC. Here, $I_{out} \times 500$ is the Analog value. For example, when the output of the upper DAC is 190 and the output of the lower DAC is 212, the output of the OEC is 80.560 162 [mA]. Since $80.560162 \times 500 = 40280.081$, the Analog value is 40280.081. On the other hand, $190 \times 212 = 40280$, so the true value is 40280. This result shows that the output of the OEC has a current error equivalent to 0.081.

The error between the Analog value and the true value helps to evaluate the accuracy of the optical circuit part of the optical components. This evaluation enables us to examine the effect of noise on the optical circuit part of the optical components and its tolerance to noise. The error between the ADC's output and the true value helps evaluate the optical component's accuracy. We name the output value of the ADC as the Digital value. We simulate Optical-AddSubAE and Optical-Round on OptiSystem as well as Optical-Multi.

4.2.2 Experiment with OEFMs

We reproduce the electrical digital components of an FP multiplier in Python to verify OEFM's arithmetic error. It is assumed that the Python-created components (electrical-digital circuits) are error-free.

4.2.3 Accuracy evaluation index for arithmetic errors

We evaluate the error tolerance of OEFM against noise by performing an accuracy evaluation. Considering various implementation situations, we set the noise variance in a wide range (10^{-15} to 10^{-3} [mW]) in the simulation, including noise larger than realistic noise, and performed sensitivity analysis. To evaluate the accuracy of the optical arithmetic units and the OEFM, we investigate the mean and standard deviation of the arithmetic error and error rate for the optical arithmetic unit and the relative error for the OEFM. X_{opt_i} is the output of the optical arithmetic units, and X_{tv_i} is the true value, each of which is an integer value. Here, the true value is the value calculated by Python on a general purpose server.

The error is X_{opt_i} minus X_{tv_i} (the difference between X_{opt_i} and X_{tv_i}) on each sample i. *mean.error* is the mean value of the arithmetic error. *mean.error* is represented by Equation (9).

$$mean.error = \frac{1}{N}\sum_{i=1}^{N}(X_{opt_i} - X_{tv_i}) \tag{9}$$

N is the number of data samples used in the experiment. In this research, N is 1000. *std.error* is the standard deviation of the arithmetic error. *std.error* is expressed by Equation (10).

$$std.error = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\{(X_{opt_i} - X_{tv_i}) - mean.error\}^2} \tag{10}$$

The error rate is defined by Equation (11).

$$Error.rate = \frac{Miss}{N} \times 100 \tag{11}$$

Miss is the number of samples of X_{opt_i} that disagree with X_{tv_i}. N is the total number of samples. The error rate is a measure of the rate of arithmetic errors.

rel.err% is the relative error of OEFMs. F_{opt_i} is the output of OEFM, and F_{tv_i} is the true value, each

of which is a floating-point value. $rel.err\%$ is defined by Equation (12).

$$rel.err\% = \frac{1}{N}\sum_{i=1}^{N}\left(\left|\frac{F_{opt_i} - F_{tv_i}}{F_{tv_i}}\right|\right) \times 100 \qquad (12)$$

4.3 Latency and energy consumption

The latency and energy consumption are estimated on the basis of the model. Latency represents the time required for calculating a set of FP multiplication inputs. The energy consumption is the energy consumed calculation of the set. The comparison is with Electrical-FM, an FP multiplier in which all components are electrical circuits.

As mentioned in Section 4.1, optical arithmetic units have a trade-off between arithmetic accuracy and latency/power consumption. Therefore, we assume that the laser light intensity is sufficient for error-free computation. Specifically, the maximum signal power of the laser is 65.025 mW and 16.129 mW for Optical-Multi and Optical-AddSubAE/Round, respectively. The typical noise variance for shot noise and thermal noise at photodetector is 10^{-11} to 10^{-10}[mW] [21], and the SNR is about 40 to 30 [dB], 80 to 70 [dB] and 85 to 75 [dB] for Optical-Multi, Optical-AddSubAE and Optical-Round, respectively. This is a large SNR compared to the SNR (30-15 [dB]) of the measured data of the fabricated optical chip [22].

We explain the model for estimating latency and energy consumption. The latency in optical circuits can be calculated using [*Path length*]/[*Speed of light in circuit*] instead of RC delay as in electrical circuits [23]. Passive optical devices do not consume energy. The energy consumption of active optical devices is 32.4 [fJ/FLO] for MZI [24] and 1.2 [fJ/FLO] for the photodetector [25]. As mentioned above, the maximum laser output is 65.025 mW for Optical-Multi and 16.129 mW for Optical-AddSubAE and Optical-Round.

The latency and energy consumption of optical components include those of ADC/DAC. The Walden model can explain the relationship between ADC/DAC latency and energy consumption [26]. For specific values, refer to the latest design values (ADC [27], DAC [28]).

The latency of the electrical circuit is calculated on the basis of the delay time of each logic gate and the Elmore delay model [29]. The energy consumption of the electrical circuit is calculated by the number of gates for each component. DSENT [30] calculates the logic gate's latency and energy consumption. A technology file (TG11LVT model) equivalent to 11 [nm] is used in DSENT. Latency and energy consumption are calculated for each component. For latency, add the latency of each component through which data passes in each of the Sign part, Exponent part, and Fraction part shown in Figure 1, and the maximum value among them will be the latency of the entire FP multiplier. Whereas, as for energy consumption, the total energy consumption of the FP multiplier is the sum of each component's energy consumption.

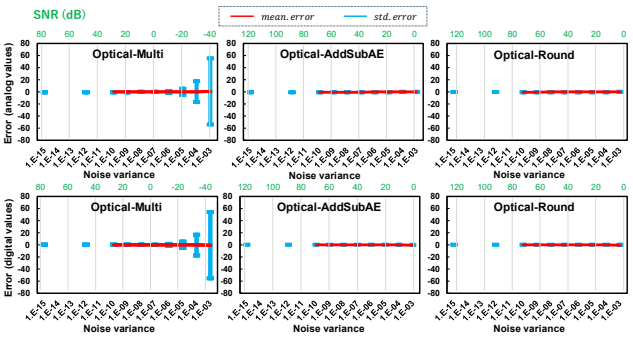


Fig. 6 Analog and digital errors for optical arithmetic units

For Electrical-FM pipelining, the throughput and energy consumption are estimated using the number of stages Nstage as a variable. Generally, circuits are sliced into stages with registers between stages, and control circuits achieve pipeline execution. In this paper, we assume ideal pipelining, ignoring control circuits and wiring, which makes it a potential comparison and analysis of optical circuits. Insert registers and divide the stages according to the following procedures.

1. For the Fraction part, we insert registers at positions that divide the latency equally into Nstages.
2. For the Exponent and Sign parts, insert the minimum registers so that the pipeline clock cycle time determined in step 1 is not exceeded.

The inserted registers are 16 bits, which is the expected maximum bit case. Pipelining of optical circuits is currently difficult due to the immaturity of memory devices. We compare and evaluate the energy consumption of pipelined electrical FM, which has the same throughput as each OEFM without pipeline technology.

5. Evaluation result

5.1 Arithmetic accuracy

Figure 6 shows the Analog and Digital errors of optical arithmetic units. The horizontal axis is the noise variance set by the Noise Source. The rightward direction represents higher noise levels. The vertical axis is the magnitude of the error. The points in the graph are the mean values of the errors, while error bars show the error standard deviation. The SNR is also shown at the top of the graph corresponding to the noise variance. For Optical-Multi, the greater the set noise variance,

Table 2 Error.rate and SNR at each noise variance

noise variance [mW]		1.00E-15	1.00E-14	1.00E-13	1.00E-12	1.00E-11	1.00E-10	1.00E-09	1.00E-08	1.00E-07	1.00E-06	1.00E-05	1.00E-04	1.00E-03
SNR (dB)	Optical-Multi	79	69	59	49	39	29	19	9	-1	-11	-21	-31	-41
	Optical-AddSubAE	120	110	100	90	80	70	60	50	40	30	21	11	0
	Optical-Round	125	115	105	95	85	75	65	55	45	35	25	15	5
Error.rate	Optical-Multi	0	-	-	0	-	0	0	0.68	34.04	76.96	91.68	97.32	99.52
	Optical-AddSubAE	0	-	-	0	-	0	0	0	0	0	0	0.4	26.16
	Optical-Round	0	-	-	0	-	0	0	0	0	0	0	0	8.08

Table 3 rel.err% at each noise variance

noise variance [mW]	1.00E-15	1.00E-14	1.00E-13	1.00E-12	1.00E-11	1.00E-10	1.00E-09	1.00E-08	1.00E-07	1.00E-06	1.00E-05	1.00E-04	1.00E-03
Ao-OEFM	0	-	-	0	-	0	0	0	1.43E-03	3.84E-03	9.05E-03	9.05E-03	9.50E-02
Lo-OEFM	0	-	-	0	-	0	0	0	1.43E-03	3.84E-03	9.05E-03	2.63E-02	1.57E+01
Eo-OEFM	0	-	-	0	-	0	0	0	1.43E-03	3.84E-03	9.05E-03	9.05E-03	1.57E+01

the larger the magnitude of the error. The Digital error becomes noticeable when the noise variance is 10^{-07} or more. For Optical-AddSubAE and Optical-Round, those components' Analog and Digital errors are about 0. Practically, the light intensity set for this evaluation is sufficiently large compared to realistic noise, and there is no error in optical arithmetic units. As mentioned in Section 4.3, the SNRs with our assumed parameters are about 40 to 30 [dB], 80 to 70 [dB] and 85 to 75 [dB] for Optical-Multi, Optical-AddSubAE and Optical-Round, respectively. Figure 6 shows that every optical unit has negligible small errors. Therefore, the parameters on which the evaluation is based are "error-free" in terms of accuracy, and the accuracy is the same as Ectrical-FM, making it a fair comparison in terms of latency and energy. Please refer to Section 5.2 for the latency and energy comparison results.

Table 2 shows the *Error.rate* of each optical component with SNR. Table 3 shows the rel.err% for the OEFMs (Ao-OEFM, Lo-OEFM, and Eo-OEFM). Optical-Multi exhibits an error when the noise variance is greater than 10^{-8}, with an error rate of 97.32% at 10^{-4}. However, the rel.err% for Ao-OEFM is not that large. This is because the error in Multi is mitigated by Normalized and Round. Since Optical-Multi performs analog arithmetic, possible errors are concentrated in the lower bits in digital; the output of Multi is 16 bits, whereas the output of Round is 8 bits, so the information in the lower bits of Multi' s output is mainly lost. Therefore, the errors that Optical-Multi has are hidden. In this respect, Optical-Multi works well with FP multipliers.

The experimental results show that the error in the exponential part (Optical-AddSubAE) significantly impacts the rel.err% of the OEFM. When the error rate of the Optical-AddSubAE is non-zero (when the noise variance is 10^{-3}), the rel.err% of the Lo-OEFM and Eo-OEFM increases significantly compared with that of the Ao-OEFM. This indicates that if accuracy is essential, it is better to perform the Exponent part electrically.

5.2 Latency and energy consumption

As shown in Figure 1, the components of the FP multiplier are classified for calculating the Sign, Exponent, and Fraction parts, respectively. Figure 7's stacked bar charts illustrate cumulative latency for each calculation, with legends corresponding to Figure 1 components. The latency of Electrical-FM is 5.51 [ns]. Figure 8 shows the cumulative energy consumption for all FP multiplier components. The unit is fJ/FLO, i.e., the energy required per FP multiplication. The energy consumption of Electrical-FM is 1327 [fJ].

The latency and energy consumption of each OEFM are described. For Ao-OEFM, by replacing Electrical-Multi with Optical-Multi, its critical path is calculated in the exponential part with a latency of 2.43 [ns], and its operating frequency is 0.41 [GHz]. The energy consumption of Ao-OEFM is 779 [fJ]. It is important to note that the higher operating frequency improves the energy consumption due to the static energy of the electrical components. Therefore, the energy consumption of the electrical components in Ao-OEFM is less than that of the same components in Electrical-FM. For example, Add's energy consumption is 74 [fJ] for Electrical-FM, while 52 [fJ] for Ao-OEFM. For Lo-OEFM, its latency is 0.68 [ns], its operating frequency is 1.48 [GHz], and its energy consumption is 926 [fJ]. For Eo-OEFM, its latency is 1.09 [ns], its operating frequency is 0.92 [GHz], and its energy consumption is 772 [fJ]. See Table 4 for the latency and energy consumption of the optical components.

Figure 9 shows the throughput and energy consumption of pipelined Electrical-FM with the number of stages as a variable. The left figure represents throughput, and the right figure represents energy consumption. Throughput improves in proportion to the number of stages. On the other hand, energy consumption increases as the number of stages increases. The energy consumption graph is not a straight line because the number of pipeline registers in the Exponent part does not match the number of stages. Note that the sign part did not need to be divided into registers.

Table 4 Latency and energy of the optical arithmetic units

Component name	Latency[ns] (Electrical)	Energy consumption[fJ] (Electrical)
Optical-Multi	0.26 (4.68)	358 (723)
Optical-AddSubAE	0.24 (2.43)	278 (364)
Optical-Round	0.26 (0.68)	279 (220)

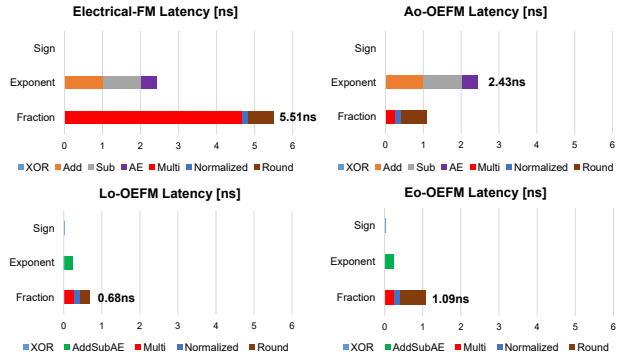


Fig. 7 Latency for OEFMs and Electrical-FM

Electrical-FM, a two-stage pipeline, achieved throughput equivalent to Ao-OEFM. Ao-OEFM consumed 43.0% less energy than Electrical-FM (two-stage pipeline). Electrical-FM, an eight-stage pipeline, achieved throughput equivalent to Lo-OEFM. Lo-OEFM consumed 41.4% less energy than Electrical-FM (eight-stage pipeline). Electrical-FM, a five-stage pipeline, achieved throughput equivalent to Eo-OEFM. Eo-OEFM consumed 48.4% less energy than Electrical-FM (five-stage pipeline). The pipelined Electrical-FM achieves throughput equivalent to OEFM ideally. However, OEFM consumes less energy than pipelined Electrical-FM. The results of this study ignore the overhead caused by the complexity of wiring, such as pipeline controllers, so the actual reduction in energy consumption is expected to be even greater.

In all cases of Ao-OEFM, Lo-OEFM, and Eo-OEFM, OEFM has lower latency and energy consumption than Electrical-FM. In particular, compared to Electrical-FM, Lo-OEFM reduces latency by 87%, and Eo-OEFM reduces energy consumption by 42%. Normalized optical implementation is effective for further performance improvement. In the Lo-OEFM Fraction calculation, the DAC/ADC latency accounts for about 59% of the total latency. On the other hand, Eo-OEFM is more energy efficient than Lo-OEFM because the energy consumption of Optical Round is larger than that of Electrical-Round due to the DAC/ADC overhead. Therefore, reducing the DAC/ADC by implementing Optical-Noramalized is effective in terms of latency and energy consumption. In this study, since the rounding mode is "round to nearest - even" and involves digital concepts, the optical implementation of Normalized is incompatible because it is an optical analog operation. An optical implementation of Normalized will be the subject of future work.

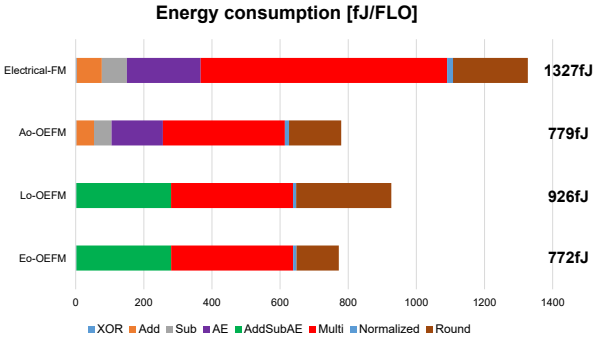


Fig. 8 Energy consumption for OEFMs and Electrical-FM

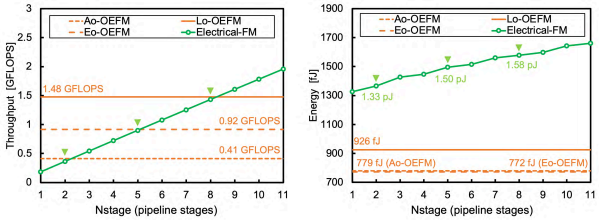


Fig. 9 Throughput and energy consumption of pipelined Electrical-FM with OEFMs

6. Conclusions

We propose three Opto-Electrical Floating-point Multipliers, which are accuracy-oriented (Ao-OEFM), latency-oriented (Lo-OEFM), and energy-oriented (Eo-OEFM), using analog optical processing to improve latency and energy efficiency. Optical devices involve a trade-off between accuracy and latency/energy consumption. Regarding arithmetic accuracy, Ao-OEFM demonstrated high noise tolerance, while Lo-OEFM and Eo-OEFM ensured sufficient accuracy. In terms of latency and energy consumption, compared to the Electrical-FM, the three OEFMs reduced latency and energy consumption, especially Lo-OEFM, achieved 87% latency reduction, and Eo-OEFM achieved 42% energy consumption reduction. By reducing converters, further latency and energy consumption reductions are expected. Developing an optical normalization implementation is a future work as it will lead to converter reduction.

Acknowledgment

This work was partially supported by JST CREST Grant Number JPMJCR21C3 and JSPS KAKENHI Grant Number JP22H05194, 22H05000, Japan.

References

[1] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle,

D. Englund and M. Soljačić: "Deep learning with coherent nanophotonic circuits", Nature photonics, **11**, 7, pp. 441–446 (2017).

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

[3] T. Inaba, T. Ono, K. Inoue and S. Kawakami: "A hybrid opto-electrical floating-point multiplier", 2022 IEEE 15th International Symposium on Embedded Multicore/Manycore Systems-on-Chip, pp. 313–320 (2022).

[4] N. Wang, J. Choi, D. Brand, C. Y. Chen and K. Gopalakrishnan: "Training deep neural networks with 8-bit floating point numbers", Advances in neural information processing systems, **31**, (2018).

[5] N. Burgess, J. Milanovic, N. Stephens, K. Monachopoulos and D. Mansell: "Bfloat16 processing for neural networks", 2019 IEEE 26th Symposium on Computer Arithmetic, pp. 88–91 (2019).

[6] S. M. Mishra, A. Tiwari, H. S. Shekhawat, P. Guha, G. Trivedi, P. Jan and Z. Nemec: "Comparison of floating-point representations for the efficient implementation of machine learning algorithms", 32nd International Conference Radioelektronika, pp. 1–6 (2022).

[7] J. O. Ríos, A. Armejach, G. Khattak, E. Petit, S. Vallecorsa and M. Casas: "Evaluating mixed-precision arithmetic for 3D generative adversarial networks to simulate high energy physics detectors", 2020 19th IEEE International Conference on Machine Learning and Applications, pp. 49–56 (2020).

[8] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam and D. Burger: "Dark silicon and the end of multicore scaling", 38th annual international symposium on Computer architecture, pp. 365–376 (2011).

[9] O. Pieters, T. De Swaef, M. Stock, S. Michiel and w. Francis: "Leveraging plant physiological dynamics using physical reservoir computing", Scientific Reports, **12**, 1, pp. 1–14 (2022).

[10] X. Fu, M. Rol, C. C. Bultink, J. van Someren, N. Khammassi, I. Ashraf, R. Vermeulen, J. De Sterke, W. Vlothuizen, R. Schouten, et al.: "A microarchitecture for a superconducting quantum processor", IEEE Micro, **38**, 3, pp. 40–47 (2018).

[11] K. Ishida, I. Byun, I. Nagaoka, K. Fukumitsu, M. Tanaka, S. Kawakami, T. Tanimoto, T. Ono, J. Kim and K. Inoue: "Superconductor computing for neural networks", IEEE Micro, **41**, 03, pp. 19–26 (2021).

[12] P. Singh, D. K. Tripathi, S. Jaiswal and H. Dixit: "All-optical logic gates: designs, classification, and comparison", Advances in Optical Technologies, **2014**, (2014).

[13] Y. Fu, X. Hu and Q. Gong: "Silicon photonic crystal all-optical logic gates", Physics letters A, **377**, 3-4, pp. 329–333 (2013).

[14] B. Dai, S. Shimizu, X. Wang and N. Wada: "Simultaneous all-optical half-adder and half-subtracter based on two semiconductor optical amplifiers", IEEE Photonics Technology Letters, **25**, 1, pp. 91–93 (2012).

[15] S. Kaur, R. S. Kaler and T. S. Kamal: "All-optical binary full adder using logic operations based on the nonlinear properties of a semiconductor optical amplifier", Journal of the Optical Society of Korea, **19**, 3, pp. 222–227 (2015).

[16] K. Kitayama, M. Notomi, M. Naruse, K. Inoue, S. Kawakami and A. Uchida: "Novel frontier of photonics for data processing—photonic accelerator", Apl Photonics, **4**, 9, p. 090901 (2019).

[17] S. AbdollahRamezani, K. Arik, A. Khavasi and Z. Kavehvash: "Analog computing using graphene-based metalines", Optics letters, **40**, 22, pp. 5239–5242 (2015).

[18] N. H. Weste and D. Harris: "CMOS VLSI design: a circuits and systems perspective", Pearson Education India (2015).

[19] O. A. L. Abdul: "Performance estimation of n-bit classified adders", International Journal of Computer Applications, **80**, 9 (2013).

[20] "OptiSystem", https://optiwave.jp/home/optisystem/.

[21] H. Zhou, W. Yang, Z. Li, X. Li and Y. Zheng: "A bootstrapped, low-noise, and high-gain photodetector for shot noise measurement", Review of Scientific Instruments, **85**, 1 (2014).

[22] L. Qiao, W. Tang and T. Chu: "32× 32 silicon electro-optic switch with built-in monitors and balanced-status units", Scientific Reports, **7**, 1, p. 42306 (2017).

[23] K. Shiflett, D. Wright, A. Karanth and A. Louri: "Pixel: Photonic neural network accelerator", 2020 IEEE International Symposium on High Performance Computer Architecture, pp. 474–487 (2020).

[24] J. Ding, R. Ji, L. Zhang and L. Yang: "Electro-optical response analysis of a 40 Gb/s silicon mach-zehnder optical modulator", Journal of lightwave technology, **31**, 14, pp. 2434–2440 (2013).

[25] K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, A. Shinya, E. Kuramochi and M. Notomi: "Forward-biased nanophotonic detector for ultralow-energy dissipation receiver", APL Photonics, **3**, 4, p. 046101 (2018).

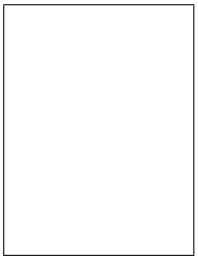
[26] R. H. Walden: "Analog-to-digital converter survey and analysis", IEEE Journal on selected areas in communications, **17**, 4, pp. 539–550 (1999).

[27] B. Murmann: "ADC performance survey 1997-2020", IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers VLSI Symp (2020).

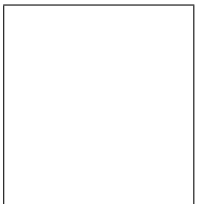
[28] X. Wu, P. Palmers and M. S. Steyaert: "A 130 nm cmos 6-bit full nyquist 3 GS/s dac", IEEE Journal of Solid-State Circuits, **43**, 11, pp. 2396–2403 (2008).

[29] A. I. Abou-Seido, B. Nowak and C. Chu: "Fitted elmore delay: a simple and accurate interconnect delay model", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, **12**, 7, pp. 691–696 (2004).

[30] C. Sun, C. O. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L. Peh and V. Stojanovic: "DSENT-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling", 2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip, pp. 201–210 (2012).

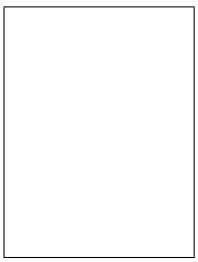


Takumi Inaba received his B.E. in the Department of Electrical Engineering and Computer Science from Kyushu University in 2020. He is currently a master's student in the Graduate School of Information Science and Electrical Engineering at Kyushu University. His research interests include nanophotonic computing and computer architecture.

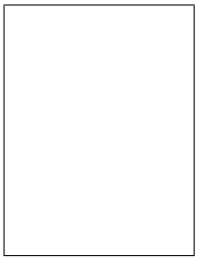


Takatsugu Ono received the Ph.D. degree from Kyushu University, Japan, in 2009. In 2010, he joined Fujitsu Labo-

ratories Ltd., Kawasaki, Japan, as a researcher. He is currently an associate professor at the Department of Advanced Information Technology at Kyushu University. His research interests include the areas of computer architecture with particular emphasis on memory systems (including non-volatile memory), secure computing, superconductor computing, nano-photonic computing, and high-performance computing. He is a member of the IEEE, the IPSJ, and the IEICE.



Koji Inoue received his B.E. and M.E. in computer science from the Kyushu Institute of Technology, Japan, in 1994 and 1996, respectively. He received his Ph.D. from the Department of Computer Science and Communication Engineering, Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan in 2001. In 1999, he joined Halo LSI Design & Technology, Inc., NY, as a circuit designer. He is currently a professor of the Department of Advanced Information Technology, Kyushu University. His research interests include superconductor computing, quantum computing, power-aware computing, high-performance computing, and microarchitecture of processor and memory systems.



Satoshi Kawakami received his B.E. and M.E. degrees in electrical engineering and computer science from Kyushu University, Japan in 2012 and 2014, respectively. After working for Bosch Corporation for two years as an engineer, he received the Ph.D. degree in information science and electrical engineering from Kyushu University, Japan in 2019. His research interests include computer system architecture with emerging technologies such as superconductors and nano-photonic devices. He is a member of IEEE, ACM, the Information Processing Society of Japan (IPSJ), and the Institute of Electronics, Information and Communication Engineers of Japan (IEICE).