Novel frontier of photonics for data processing —Photonic accelerator

Kitayama, Ken-ichi
Department of Optical Information and Systems, Graduate School for the Creation of New Photonics Industries

Notomi, Masaya
NTT Basic Research Laboratories

Naruse, Makoto
Department of Information Physics and Computing, The University of Tokyo

Inoue, Koji
Department of Informatics, Kyushu University

他

出版情報：APL Photonics. 4 (9), pp.090901-, 2019-09-24. AIP Publishing
バージョン：

APL Photonics

Novel frontier of photonics for data processing—Photonic accelerator

Cite as: APL Photonics 4, 090901 (2019); https://doi.org/10.1063/1.5108912
Submitted: 03 May 2019 . Accepted: 18 August 2019 . Published Online: 24 September 2019

Ken-ichi Kitayama, Masaya Notomi, Makoto Naruse, Koji Inoue, Satoshi Kawakami, and Atsushi Uchida

Novel frontier of photonics for data processing—Photonic accelerator

Cite as: APL Photon. 4, 090901 (2019); doi: 10.1063/1.5108912
Submitted: 3 May 2019 • Accepted: 18 August 2019 •
Published Online: 24 September 2019

Ken-ichi Kitayama,[1,2] Masaya Notomi,[3] Makoto Naruse,[4] Koji Inoue,[5] Satoshi Kawakami,[5] and Atsushi Uchida[6]

AFFILIATIONS

[1]Department of Optical Information and Systems, Graduate School for the Creation of New Photonics Industries, Hamamatsu 431-1202, Japan
[2]Network System Research Institute, National Institute of Information and Communications Technology, Koganei 184-8795, Japan
[3]NTT Basic Research Laboratories, Atsugi 243-0198, Japan
[4]Department of Information Physics and Computing, The University of Tokyo, Tokyo 113-8656, Japan
[5]Department of Informatics, Kyushu University, Fukuoka 819-0395, Japan
[6]Department of Information and Computer Sciences, Saitama University, Saitama 338-8570, Japan

ABSTRACT

In the emerging Internet of things cyber-physical system-embedded society, big data analytics needs huge computing capability with better energy efficiency. Coming to the end of Moore's law of the electronic integrated circuit and facing the throughput limitation in parallel processing governed by Amdahl's law, there is a strong motivation behind exploring a novel frontier of data processing in post-Moore era. Optical fiber transmissions have been making a remarkable advance over the last three decades. A record aggregated transmission capacity of the wavelength division multiplexing system per a single-mode fiber has reached 115 Tbit/s over 240 km. It is time to turn our attention to data processing by photons from the data transport by photons. A photonic accelerator (PAXEL) is a special class of processor placed at the front end of a digital computer, which is optimized to perform a specific function but does so faster with less power consumption than an electronic general-purpose processor. It can process images or time-serial data either in an analog or digital fashion on a real-time basis. Having had maturing manufacturing technology of optoelectronic devices and a diverse array of computing architectures at hand, prototyping PAXEL becomes feasible by leveraging on, e.g., cutting-edge miniature and power-efficient nanostructured silicon photonic devices. In this article, first the bottleneck and the paradigm shift of digital computing are reviewed. Next, we review an array of PAXEL architectures and applications, including artificial neural networks, reservoir computing, pass-gate logic, decision making, and compressed sensing. We assess the potential advantages and challenges for each of these PAXEL approaches to highlight the scope for future work toward practical implementation.

I. INTRODUCTION

We create nearly 2.5 exabytes (10^{18}) of data every day, and surprisingly, this information amounts to about a half of the information of the human genome on this planet. This exponential increase of daily data generation in the world today has led to a new era in data exploration and utilization. In the emerging Internet of Things (IoT)-cyber physical system (CPS)-embedded society,[1] big data analytics need a huge computing capability with better energy efficiency. The way people use computers has been radically changing to cloud computing without the ownership of a computing resource. Cloud computing provides the users with various types of services, the so-called X as a service (XaaS), where X stands for software (S), platform (P), and desktop (D). On the other hand, for time-sensitive applications such as autonomous vehicle and augmented reality/virtual reality (AR/VR) and industrial robots, edge/fog computing[2] in a microdirectional couplers (DCs)[3] can be provided with its computing resource to a site as close as possible where the events happen.

Now the computing resource has become available anywhere via mobile computing, e.g., over a 5G wireless link[4] between a smart tablet and cloud or edge/fog.

Let us turn our attention to computing hardware. Being underpinned by Dennard scaling,[5,6] device integration in electronic circuits such as microprocessor chips has been making a steady progress at the pace of Moore's law,[7] that is, the number of transistors incorporated in a chip approximately doubles every 18–24 months and the performance per watt grows at this same rate. As Moore's law has been coming to the end, the clock frequency of processor levels off after 2004.[8] Then, multiple processors help sustain a steady growth of the throughput by parallel computation at the expense of energy consumption. According to Amdahl's law,[9] however, the speedup of parallel computation is eventually limited, and the parallel computation cannot resolve all the issues after the end of Moore's law.

As for approaches to the resolution from the integration of devices, various post-Moore optoelectronic circuit technologies are now being developed under guidelines referred to as "More Moore," "More than Moore," and "Beyond complementary metal–oxide–semiconductor (CMOS)."[10,11] An approach of More than Moore combined with computing architectures provides a path to hardware accelerators. The hardware accelerator is a special class of processors such as the graphics processing unit (GPU) and field-programmable gate array (FPGA) placed at the front end of digital computer, which is optimized to perform a specific function.

Optical fiber transmissions have been making a remarkable advance over the last three decades. A record aggregated transmission capacity of wavelength division multiplexing (WDM) system per a single-mode fiber has reached 115 Tbit/s over 240 km.[12] Furthermore, another record of 10.16-peta-bit/s space division multiplexing (SDM)/WDM transmission using a 6-mode 19-core fiber over 11.3 km has been demonstrated.[13] It is time to turn our attention to data processing by photons from the data transport by photons. The photonic accelerator (PAXEL) is a special class of processors placed at the front end of a digital computer, which is optimized to perform a specific function but does so faster with less power consumption than an electronic general-purpose processor. It can process images or time-serial data either in the analog or digital fashion on a real-time basis. Having had the maturing manufacturing technology of optoelectronic devices and a diverse array of computing architectures at hand, prototyping PAXEL becomes feasible by leveraging on, e.g., cutting-edge miniature and power-efficient nanostructured silicon (Si) photonic devices.

In this article, first the bottleneck and the paradigm shift of digital computing are reviewed. Next, we review an array of PAXEL architecture and applications, including artificial neural networks (ANNs), reservoir computing, pass-gate logic, decision making, and compressed sensing. We assess the potential advantages and challenges for each of these PAXEL approaches to highlight the scope for future work toward practical implementation.

II. PARADIGM SHIFT OF DIGITAL COMPUTING

A. Bottleneck of general digital computing and their optoelectronic solutions

Discussion hereafter is visually summarized in Fig. 1 for quick understanding. Dennard scaling is an engineering rule of the

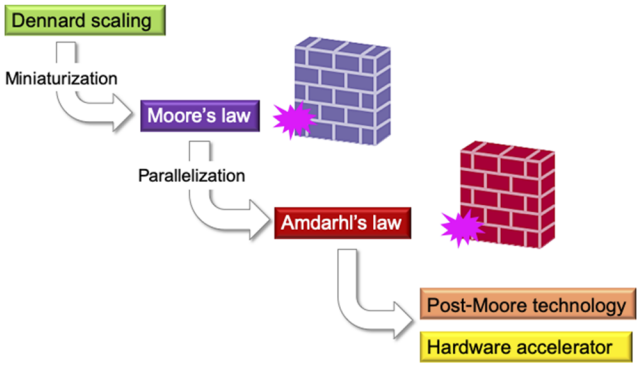

FIG. 1. Evolution and bottleneck of electronic integrated circuit for digital computing.

metal–oxide–semiconductor field-effect transistor (MOSFET) that the smaller the transistor becomes, the better the performance.[5,6] The scaling rule states that as the gate length and width of the transistors shrink by a factor of $1/k$, the circuits could operate at k-times higher frequency with smaller power by a factor of $1/k^2$. Being underpinned by the Dennard scaling, the device integration of electronic circuits such as microprocessor chips had been making a steady progress. The number of transistors incorporated in a chip approximately doubles every 18–24 months, and the performance per watt grows at this same rate. This is known as Moore's law.[7] This trend is well supported by Koomey's extensive survey on commercial products of CPUs that the number of computations per joule has been doubling approximately every 1.57 years for over 40 years until the year of 2000.[10] However, the clock frequency of processors leveled off after around 2004,[8] and Barr's analysis also evidences that after 2000, the doubling period had slowed down to every 2.6 years as shown in Fig. 2.[11]

Upon the end of Moore's law for the complementary CMOS, a multiple processor helps sustaining a steady growth of the throughput in parallel computation. Parallel computation cannot resolve all the issues; however, according to Amdahl's law,[9] the speed-up S, which is limited by the serial part of the program, is expressed by

$$S = \frac{1}{(1-p) + \frac{p}{u}}, \tag{1}$$

$$\lim_{s\to\infty} S = \frac{1}{1-p}, \tag{2}$$

where u is the number of processors and p denotes the proportion of the execution time that the part benefiting from the parallelization. In Fig. 3, the speed-up S is numerically shown for $p = 0.95$, 0.90, 0.75, and 0.50. For example, as indicated by the curve in blue, if a program needs 20 h using a single processor and a particular part of the program which takes, say, 1 h to execute cannot be parallelized, regardless of how many processors are devoted to the parallelized execution of this program, while the remaining 19 h ($p = 0.95$) of execution time can be parallelized, then the minimum execution time cannot be less than that critical 1 h. Hence, from Eq. (2), the theoretical speed-up is limited to at most 20 times ($\frac{1}{1-p} = 20$). For this reason, parallel computing with many processors is useful only for highly parallelizable programs.

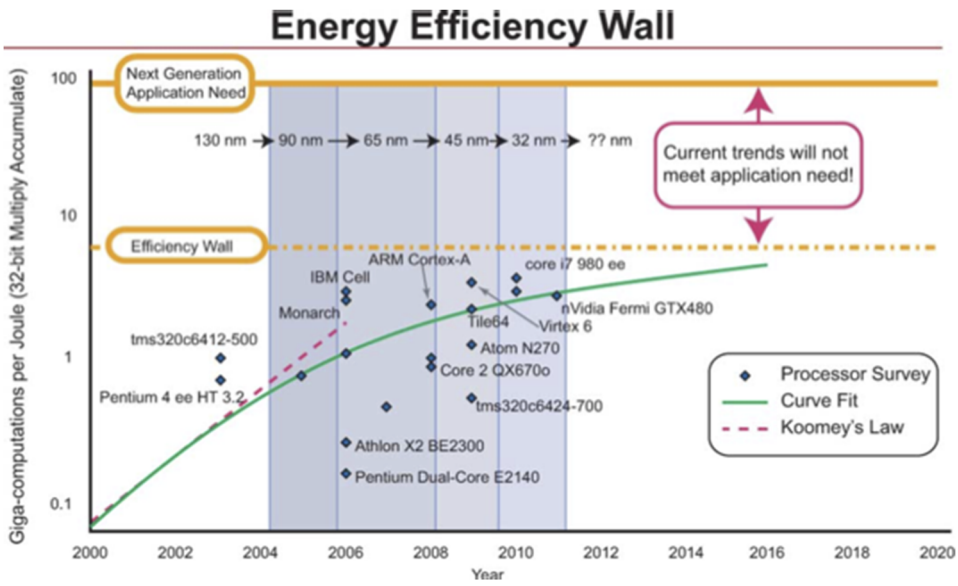

FIG. 2. Energy efficiency (giga-multiply-accumulations per Joule) vs year for commercial digital processors. Note that Koomey's law, which assumes operations per energy increases at a faster rate than the Moore's law rate of 2× per every 1.5 years no longer holds as of about 2005 and operations per unit energy starts to fall off creating a larger gap between practice and theory. Reprinted with permission from Marr *et al.*, IEEE Trans. Very Large Scale Integr. Syst. **21**(1), 147–151 (2013). Copyright 2013 IEEE.

Approaches to the resolution from the integration in electronic/optoelectronic circuits, the so-called post-Moore technologies are now being developed under guidelines referred to as "More Moore," "More than Moore," and "Beyond CMOS."[14,15] A brief introduction to these three post-Moore technologies are as follows:

1. More Moore: Extremely scaled CMOS logic and memory devices

A fundamental issue that limits physical scaling of MOSFETs is quantum mechanical tunneling, which dramatically increases the off leakage current. This is because as the length of the gate channel becomes thinner, the number of energy levels decreases and eventually disappears. For typical parameters of Si FET, this estimate results in a minimum channel length of around 4 nm. Toward the densest arrangement, at least three-dimensional (3D) packaging will play an increased role, as there is no visible alternative seen that could offer further improvements in scaling.

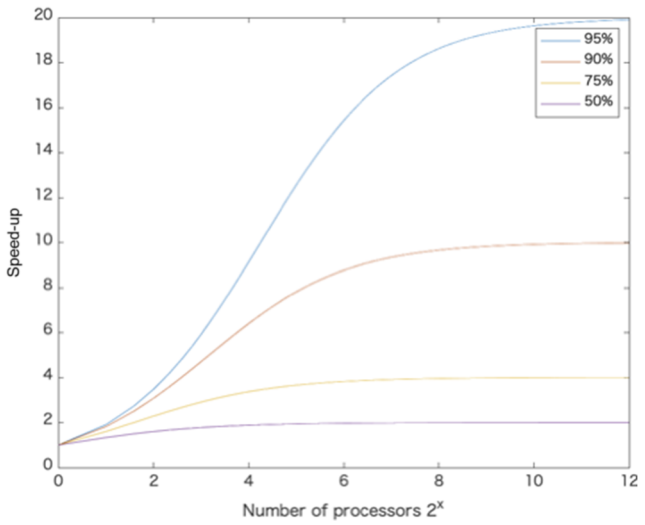

FIG. 3. Latency of the execution of a program as a function of the number of processors.

2. More than Moore: Extension of functionality of CMOS circuits by integration with other technologies

For instance, a multifunctional combination within Si technology is provided by the integration of microelectromechanical system (MEMS) devices with CMOS. A more ambitious path will not just be the integration of different functions within one material system but also the integration of different technologies, for instance, Si and III–V semiconductors. A successful example is an integrated wavelength division multiplexing (WDM) receiver chip, consisting of arrayed waveguide-grating (AWG) multichannel wavelength filters and germanium photodiodes (PDs) array using through-silicon via (TSV) technology. The area of 100-ch WDM receiver chip is reduced to 1 cm^2, which is 1/100 smaller than the conventional one.[15] Si or Si-compatible optical components have been gaining attention by copackaging with CMOS circuits, although attempts to obtain efficient Si-based light sources (LCs) have instead not been very successful.

There would be a common consensus that photonics Moore's law would never be the case because there is no scaling rule in the photonic integrated circuit (PIC) equivalent to Dennard's, which underpins Moore's law. There is no advantage seen by miniaturizing the PIC device size. The fundamental nonlinearity of a Si crystal is third order because the inversion of the symmetry of a nonstrained Si crystal prohibits the existence of the second-order nonlinearity. The third-order nonlinear coefficient is relatively small by a factor of 10^{-4} by comparing, say, the InGaAs/InAlGaAs multiquantum well (MQW). As a consequence, the nonlinear effect cannot be an obstacle under normal operating conditions with the input power less than 10 mW. It is noteworthy that in the future, the loss of Si

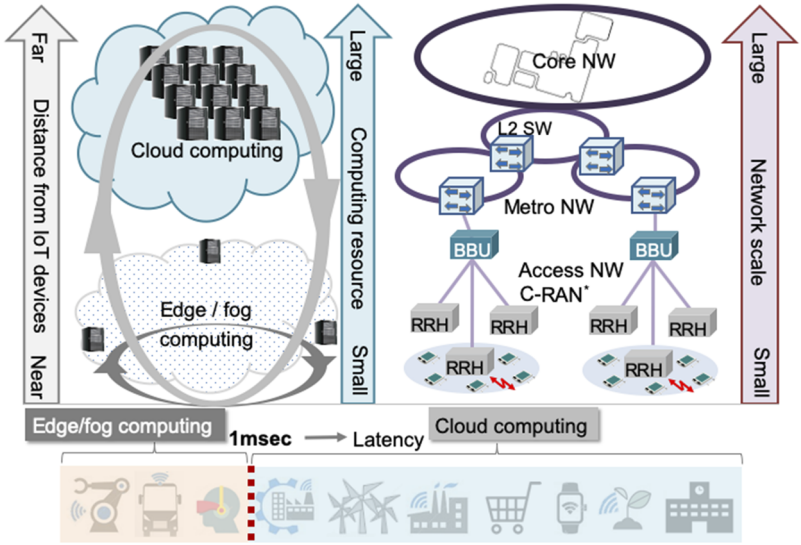

FIG. 4. From cloud computing to edge/fog computing. BBU and RRH are the baseband unit and the remote radio head of the centralized-radio access network (C-RAN) of 5G mobile communication.

waveguide will be reduced below 2 dB/cm and the third-order nonlinearity will not be negligible for the devices longer than 1 cm.

3. Beyond Moore: Alternative to CMOS transistor

III–V compound semiconductors and Ge FETs are considered viable alternatives to extend CMOS to the end of the roadmap.[16] Nanowire FETs are another candidate, in which the conventional planar MOSFET channel is replaced with a semiconductor nanowire. Recently, integrated photonics is regarded as an enabling technology for a wide range of areas. The American Institute for Manufacturing Integrated Photonics (AIM Photonics), an initiative of U.S. Federal Government, has been launched in 2015 to establish manufacturing technology of PIC by developing a widely accepted set of processes and protocols for the design, manufacture, and integration of photonics systems.[17] However, Si photonics has its limit to the miniaturization due to the weak photon energy confinement, as the size of the waveguide becomes thinner. Challenges include overcoming severe fabrication tolerance of the device geometry such as the thickness and width of the Si waveguide, large nonlinear effects of about 100 times as large as silica, and poor carrier mobility. It would be worthy to mention that photonics Moore's law can never be the case because there is no scaling rule in PIC equivalent to Dennard's. By down-sizing the PIC, the losses due to scattering caused by the roughness of the surface and bending, and light-in and light-out are much incurred. The nonlinear effect of Si due to third-order nonlinearity cannot be an obstacle under normal operating conditions with the optical power less than 10 mW. However, when the loss of the Si waveguide is reduced below 2 dB/cm in the future, the nonlinearity will not be negligible for devices longer than 1 cm.

Despite various limits which we are facing at the end of Moore's law, a possible solution path to maintain sustainable growth of digital computing capability would be a hardware accelerator, which is not almighty but optimized to perform a specific task. Electronic accelerators such as GPU, TPU (Tensor Processing Unit), and FPGA have been widely used; however, it would be better to have as many options of accelerators as possible to cope with any demands. In Sec. III, we will present photonic accelerators which will fill the gap of demands that electronic counterparts cannot meet.

B. From cloud to edge/fog computing

The way people use computers have been changing to cloud computing without the ownership of the computing resource. Cloud computing is a virtualized computing platform, which provides various types of services, including infrastructure as a service (IaaS), software (SaaS), platform (PaaS), and desktop (DaaS). It typically takes 150–200 ms of latency for data transport from sensors all the way to cloud computing in hyperscale DCs in a core network. Time-sensitive applications such as autonomous vehicle and augmented reality/virtual reality (AR/VR) and industrial robots cannot tolerate such slow responses but they require ultralow latency below 1 ms. For the time-sensitive applications, therefore, edge/fog computing would serve better, providing its computing resource in distributed micro-DCs on the network edge where the events happen. The reason why it is referred to as fog[2] is because it resembles a thinner cloud closer to the ground, meaning that it is a smaller computing resource than cloud located closer to the event, as shown by the arrow in the middle in Fig. 4. The computing resource in microdatacenters distributed around the network edge will be a key enabler to realize mobile computing by accessing it from a smart tablet through a 5G wireless link. One can carry only a light-weight tablet with a microsensing module for data acquisition and data cleansing, while edge/fog computing provides intelligence for the data processing on a real-time basis.

III. PHOTONIC ACCELERATORS: CHARACTERISTICS AND CHALLENGES

A. Hardware accelerator

The aforementioned post-Moore technologies in Sec. II A are approaches from devices to find a pathway to resolve the bottleneck of digital computing. A hardware accelerator is another approach

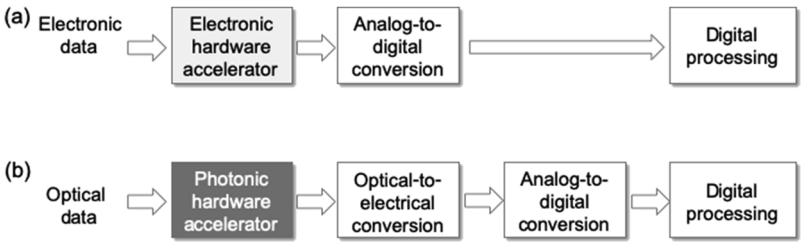

FIG. 5. (a) Electronic accelerator and (b) photonic accelerator.

to overcome the bottleneck. The hardware accelerator is a special class of processor placed at the front end of a digital computer, which is optimized to perform a specific function but does the process faster with less power consumption than an electronic general-purpose processor. Aside from conventional electronic accelerators in Fig. 5(a), our focus is on the photonic accelerator, which can directly process the data before the optical-to-electrical (OE) conversion in Fig. 5(b). Photonic accelerator will find a niche by optimizing the functionality to a specific input as is described in Sec. III. It is worthy to note that, to the authors' knowledge, the terminology of the photonic accelerator first appears in Ref. 18 whose function is based upon the time-stretch technique to capture ultrafast phenomena by a single shot, but we will use this terminology in a broader sense so that a wide variety of computing paradigms relevant to photonic technology is included.

One of the envisioned use cases of the photonic accelerator in the IoT-CPS era would be a mobile sensing and processing system, in which a compact optical sensing module is attached onto a smart tablet, as shown in Fig. 6. The compact module performs sensing and data cleansing, and then, the data are transported via a 5G wireless link, such as ultrareliable ultralow latency communication (URLLC),[19] and processed by edge/fog computing. The data acquisition and cleansing could be accelerated by the PAXEL, and thus, micro-DC benefits from off-loading the task to reduce the power consumption. A use case with a handy lens-free holographic microscopy, consisting of a single-pixel photodetector and a LED light source, will demonstrate how the photonic interface captures the data and the captured in-line holograms of the target object can be computationally reconstructed to a high-resolution image.[20] This demonstrates how the data cleansing is conducted on-site, in contrast to postprocessing of distorted image data.

Hereafter, we will focus on possible PAXEL approaches, namely, ANNs, reservoir computing, pass gate logic, decision making machines, and compressed sampling (CS).

B. Artificial neural network accelerator

Since power consumption is a primary concern in modern computer system designs, improving power efficiency, i.e., maximizing OPS/W (operations per second per watt), on neural network executions is critical. To tackle this issue, researchers have so far proposed electronic neural network accelerators like DaDianNao which is an application specific processor targeting convolutional neural networks and can achieve significant improvement in power efficiency, compared with traditional computing platforms such as CPUs and GPUs.[21] Another representative implementation is the TPU which is widely used for accelerating commercial AI services.[22] The most notable feature of such accelerators is to equip with a large number of MACs (multiply-accumulate circuits) in order to calculate vector-by-matrix multiplications (VMMs) efficiently; for instance, the first generation of TPU contains 8-bit 64K MACs in a chip. From the viewpoint of neural network applications, on the other hand, another interesting characteristic is a good tolerance to errors in computing results. This feature makes it possible to apply analog processing to aggressively improve the power efficiency, e.g., ISAAC proposed in Ref. 23 exploits analog arithmetic in crossbars.

Artificial neural network (ANN) accelerators exploiting nanophotonic devices are promising for the following reasons. First, the VMMs can be implemented by using optical elements, and we can expect ultralow latency operations because the computation is done as the light propagates in the nanophotonic device. Electronic digital operations fundamentally require charging and discharging of capacitors, and such a mechanism consumes a large amount of electric power. Since the operation speed or the clock frequency in state-of-the-art electronic digital systems is limited by the power dissipation that does not scale-down with the transistor shrinking anymore, we cannot expect improving the clock frequency. Unlike such traditional electronic digital circuits, nanophotonic devices can

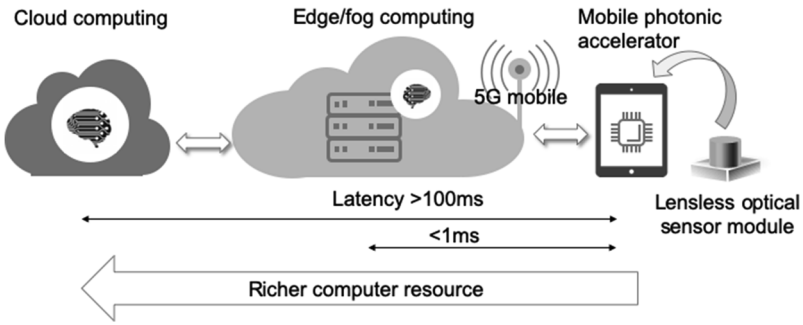

FIG. 6. Mobile hardware accelerator along with edge/fog computing via a 5G wireless link.

operate at the speed of light in the analog fashion, and such a feature makes it possible to design an ultralow latency computing platform. Although one of the critical issues of optical analog circuits is the noise, a good error tolerance of neural network applications has a significant potential to mitigate such disadvantages. Second, optical processing of high parallelism, inherent to neural network operations, is enabled by taking advantage of various attributes of light waves such as the wavelength, phase, polarization, and amplitude. Third, the high I/O bandwidth provided by an optical interconnect is important even in neural network accelerations. It has been discussed that such a nanophotonic neural network accelerator has the potential to achieve at least three orders of magnitude higher efficiency compared with an ideal electric computer.[24] Designing VMM-optimized hardware is now entering the mainstream, and this tendency strongly supports that photonic ANN research could have a real-world impact toward energy efficient AI computing.

1. Optical analog vector-by-matrix multiplier

We describe optical analog VMMs. The matrix product is the basic operation of many kinds of information processing, especially approximate computing such as a neural network. VMM requires generally a long delay in CMOS-based circuits, and hence, a significant acceleration can be expected by applying optical VMM. The operation of VMM is defined by

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}, \tag{3}$$

where $\boldsymbol{x}$ is the input vector, $\boldsymbol{W}$ is the matrix, and $\boldsymbol{y}$ is the output vector. Figure 7 illustrates three types of optical VMM implementations for the case with $N = 4$. The latency of each VMM against the number of pixels ($N \times N$) is summarized in Table I. The latency is determined by the longest optical path for the light emitted from the light source (LS) to the photodetector (PD). The longest optical path of $N \times N$ VMM mainly depends on the number of modulators on the longest pass and the waveguide length. The number of modulators as light wave passes from LSs to PDs depends on VMM configurations, as shown in Table I. The waveguide length is N, which is constant regardless of the VMM configuration. Therefore, the latency of each VMM in Table I is calculated by adding these two factors. For example, the latency of spatial light modulator (SLM)-VMM is $\mathcal{O}(N)$ since the sum of two factors is $1 + N$. As a consequence, there is no significant difference in performance depending on the size of VMM. Another important consideration is the energy consumption. When an optical signal passes an optical modulator, it loses a small amount of energy. This means that the amount of lost energy is roughly proportional to the number of optical devices, i.e., the first row of Table I. If it is assumed that a constant amount of energy is lost in each optical modulator, the SLM-VMM consumes the lowest energy.

a. SLM-VMM. One configuration is a spatial light modulator (SLM) VMM as shown in Fig. 7(a).[25] In the SLM-VMM, the input vector $\boldsymbol{x}$ is represented by a set of light sources, the weight matrix $\boldsymbol{W}$ by an SLM, and the output vector $\boldsymbol{y}$ by a set of photodetectors. SLM-VMM is scalable because it utilizes free space optics. On the other hand, a planar integrated SLM-VMM is compact and compatible with electronic VLSI technologies and microfabrication.[26] Unlike traditional electrical circuits that exploit voltage levels in order to represent digitalized information, data values treated in the SLM-VMM are mapped onto the levels of light intensity in analog form. Multiplications and additions are performed by weakening the light intensity and collecting multiple light waves at a photodetector, respectively. Since the intensity of a light source I_{in} is reduced to I_{out} after passing an SLM cell, the light transfer characteristic can be expressed as $I_{out} = \alpha \times I_{in}$, where α is a coefficient regarding the transmittance of the SLM cell. Therefore, the output element $y_1 = w_{11} \times x_1 + \cdots + w_{14} \times x_4$ depicted in Fig. 7(a) can optically be obtained by associating I_{in} and α with $\boldsymbol{x}$ and $\boldsymbol{W}$, respectively, and by gathering the weakened light sources corresponding to the first row of the matrix at the first photodetector. Another method to implement the SLM is to use an optomechanical micromirror device based on microelectromechanical system (MEMS) technology. This

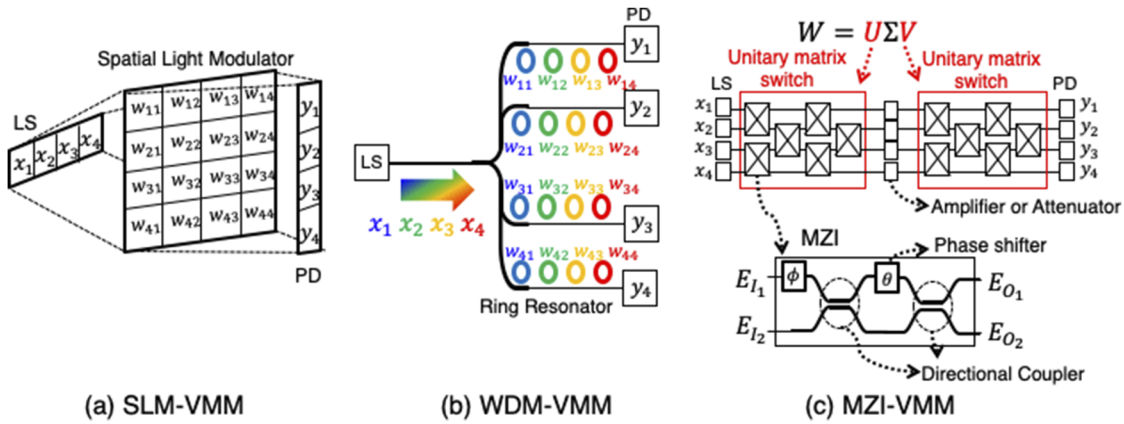

FIG. 7. Optical vector by matrix multipliers (VMMs). (a) SLM-VMM. Adapted from Ref. 26. (b) WDM-VMM. Adapted from Tait *et al.*, Sci. Rep. **7**(1), 7430 (2017). Copyright 2017 Author(s), licensed under a Creative Commons Attribution 4.0 License. (c) MZI-VMM. Adapted with permission from Shen *et al.*, Nat. Photonics **11**(7), 441–446 (2017). Copyright 2017 Springer Nature.

TABLE I. Optical circuit's latency for $N \times N$ (pixel) VMMs.

	SLM-VMM	WDM-VMM	MZI-VMM
The number of modulators (on the longest pass)	1	N	$(2N-3) \times 2$ (Reck's circuit) $N \times 2$ (Clements's circuit)
Waveguide length	N	N	N
Latency	N	$2N$	$5N$ (Reck's circuit) $3N$ (Clements's circuit)

device deflects light signals digitally by selectively redirecting parts of them.

b. WDM-VMM. Another configuration is a wavelength division multiplexing (WDM) VMM[27] as shown in Fig. 7(b). In WDM-VMM, the input vector $\boldsymbol{x}$ is represented by a wavelength multiplexed light from light sources, the matrix $\boldsymbol{W}$ by cascaded ring resonators, and the output vector $\boldsymbol{y}$ by a set of photodetectors. The WDM input signal, in which each element of the input vector is assigned a unique wavelength carrier, is equally split so that it passes through a series of ring resonators. The output is then summed up by the photodetectors. Ring resonators are placed next to a waveguide to couple only the light of a specific wavelength. When the length of the ring circumference is equal to an integer number of wavelengths which is the resonance condition, the optical signal in the waveguide is dropped to the ring and lost due to the scattering. The levels of light intensity are considered as data values in analog form as well as SLM-VMM explained above, and multiplications are implemented by the ring resonators in which the loss rate can be controlled by injecting a charge into the ring or changing its temperature. For instance, if we assume $(x_1, x_2, x_3, x_4) = (1.0, 1.0, 1.0, 1.0)$ and $(w_{11}, w_{12}, w_{13}, w_{14}) = (0.0, 1.0, 1.0, 1.0)$ in Fig. 7(b), the intensity level of x_1 becomes zero due to the effect of the associated ring resonator, and the other signals regarding x_2, x_3, x_4 are gathered to obtain the final result of y_1 by the associated photodetector. In general, cascading multiple microring resonators can cause bandwidth narrowing and increase the control operation complexity. Hence, selecting appropriate parameters and well-design for microring resonators is quite important by exploiting design space like in Ref. 29. Recently, high-resolution optical WDM-VMM has also been demonstrated.[28] The number of microring resonators is the same as that in Ref. 27, but the topology is different. This work uses microring resonators with a low quality (Q) factor and demonstrates high linearity and high resolution.

c. MZI-VMM. Another configuration is a Mach-Zehnder interferometer (MZI) based VMM[24] as shown in Fig. 7(c). The MZI consists of two directional couplers (DCs) where two input signals are equally divided into two output ports and two phase shifters (PSs) that modulate the input signals. The amount of phase shift (φ, θ) can be controlled by charging the phase shifters or changing their temperature. The transfer function of MZI can be expressed by

$$\begin{pmatrix} E_{O_1} \\ E_{O_2} \end{pmatrix} = e^{i\left(\frac{\theta+\pi}{2}\right)} \begin{pmatrix} e^{i\phi}\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \\ e^{i\phi}\cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \end{pmatrix} \begin{pmatrix} E_{I_1} \\ E_{I_2} \end{pmatrix}, \tag{4}$$

where E_{O_1}, E_{O_2} and E_{I_1}, E_{I_2} are electric field amplitudes of the output and input signals, respectively. The MZI can distribute the light signal of each input port to the two output ports at an arbitrary ratio by adjusting (φ, θ) so that it can be used to implement a 2×2 unitary transformation. Reck *et al.* and Clements *et al.*[30,31] have proposed that arbitrary $N \times N$ unitary transformation can be realized by exploiting the property of MZI. As shown in Fig. 7(c), the matrix of the VMM is represented by two MZI-based unitary matrix switches and a set of either attenuators or amplifiers. The unitary matrix switches representing the $M \times M$ matrix U and $N \times N$ matrix V perform together a couple of unitary conversions. The attenuators or amplifiers represent an $M \times N$ diagonal matrix Σ of singular value decomposition (SVD). Note that Clements's implementation has features of higher fidelity, excellent tolerance to noise, and a smaller size compared with Reck's one. These differences are due to the arrangement of the MZIs: the unitary matrix switches are triangular in Reck's circuit but rectangular in Clements's implementation. Detailed operation principles and mathematical proofs can be seen in Refs. 30 and 31.

2. Power and performance modeling

Maximizing power efficiency of optical analog VMMs is a critical challenge on photonic ANN accelerators. Here, we focus on the Clements's MZI-VMM implementation in Table I and define the throughput or performance as the number of multiply accumulate operations per second. It can be evaluated for the VMM scale with the $N \times N$ matrix by

$$Throughput = N^2 \times f_{VMM}, \tag{5}$$

$$f_{VMM} = \min(f_{LS}, f_{PD}, f_{PS}, f_{DPATH}), \tag{6}$$

$$f_{DPATH} = 1/L = 1/\left(3N \times S_{MZI} \times \frac{refrective\ index}{c}\right), \tag{7}$$

where f_{VMM} is the operation frequency. f_{VMM} is limited by several factors such as the maximum operation frequency of light sources, f_{LS}, that of photodetectors, f_{PD}, two phase shifters in the MZI, f_{PS}, and that of the optical data-path f_{DPATH}, which is the reciprocal of the VMM circuit latency L. In the MZI-VMM, L depends on the length of the longest path from a light source input to a photodetector output in the optical circuit and can be modeled as Eq. (7), where S_{MZI} is the size (or length) of an MZI. From the viewpoint of photonic neural-network executions, f_{PS} and f_{DPATH} directly limit the maximum operation frequency to feed the next input data regarding $\boldsymbol{W}$ and $\boldsymbol{x}$, respectively.[32]

The energy cost is mainly incurred by the phase shifters and photodetectors. A light signal from a light source is detected by the photodetector at the loss of energy by passing through phase shifters, and then, it is converted into the photocurrent. Assume that all the energy of the optical signal at the photodetector is converted (in fact, conversion efficiency <30%) irrespective of the values of $\boldsymbol{W}$ and $\boldsymbol{x}$. The power consumption of the VMM is expressed by

$$Power = N \times P_{LS}, \tag{8}$$

where P_{LS} is the power of the light source. This power model can be applied to closed systems in which only light sources can provide energy for optical calculations, and final outputs are detected only by photodetectors. Based on Eqs. (5) and (8), the power efficiency of the optical VMM is derived as follows:

$$Power\ efficiency = \frac{N \times f_{VMM}}{P_{LS}}. \tag{9}$$

3. Impact of device/architecture/application-level codesign for power efficient optical VMM

Owing to maturing nanophotonic device technology, ANN acceleration has a good prospect for low-power, low-latency, and high-throughput computing; however, there are some drawbacks arising from the optical devices. Unlike traditional electronic digital computers, for instance, optical analog calculations are prone to suffer from the noise, resulting in degradation of computing accuracy. Another factor is the insertion loss of the phase shifter. To maximize the power-performance potential, it is required to take full advantage of nanophotonic devices and at the same time to circumvent the shortcomings. Cross-layer interactions, or device/architecture/application-level codesigns, are crucial to resolve these issues. There is a trade-off between the power efficiency and computation accuracy in optical VMMs. If target applications are tolerant to computational errors like neural network inference, a drastic power reduction can be achieved, as described hereafter. Device designers tend to concentrate only on minimizing the device footprint, and hence, sometimes, little attention has been paid to how far such a codesign optimization has an impact on system-wide power efficiency.

First, we introduce our evaluation platform as shown in Fig. 8.[32] The evaluation platform uses the hardware configurations as the inputs, including microarchitectural parameters such as the size of VMM N, device parameters such as the MZI size, and some of noise parameters. Some of these parameters are interdependent. For example, there is a trade-off between the transmittance or loss of MZI and the modulation bandwidth of PS. However, in this evaluation, these parameters are assumed to be independent to clarify which parameter contributes to the power efficiency. In other words, we ignore the dependence of parameter values in order to indicate the direction of device parameter improvement. Note that the power loss of MZI is defined as $loss = -10 \times \log_{10}(transmittance)$, and therefore, the evaluation of the power efficiency using the transmittance is equivalent to the evaluation considering the power loss.

In the software configuration, users can set neural network parameters such as the network structure, hyperparameters, and the dataset for machine learning. The platform includes an optical simulation engine and power-performance models.

Here, we discuss the impact of the cross-layer codesign on Clements's MZI-VMM circuit in Table I. The blue hatched part of the evaluation platform in Fig. 8 is used in this evaluation. Table II shows some of the representative values of the design parameters we focus on, and we expect the performance improvement for MZI-VMM from "Standard" to "Advanced" in the near future. Note that in this evaluation, the calculation accuracy depends on the effects of the shot noise of the light source, the shot noise of the dark current and the thermal noise of PD, and the fluctuation of modulators. Each of them has been obtained based on general parameter values, $\sigma^2_{shot} = 7.02 \times 10^{-11}$, $\sigma^2_{dark\_current} = 1.60 \times 10^{-17}$, $\sigma^2_{thermal} = 1.66 \times 10^{-12}$, and $\sigma^2_{mod} = 10^{-15}$. We assumed that the power of all the light sources ranges from −40 dBm to 20 dBm. It is also assumed that the signal calculated by the analog VMM is quantized by 8 bits, and the maximum operation frequency of light sources f_{LS} is the same as that of photodetectors f_{PD}.

Figure 9 shows our evaluation results. All results are normalized to the power efficiency obtained by the design using typical values

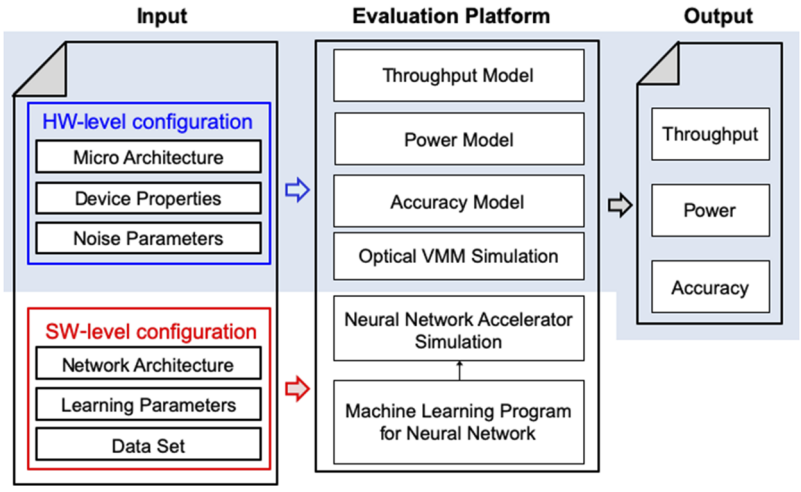

FIG. 8. Evaluation platform for nanophotonic neural-network acceleration.

TABLE II. Variable parameters.

Symbol	Description	Standard	Advanced
S_{MZI} (μm)	Size (or length) of MZI	100	1
f_{PS} (GHz)	Frequency of PS	12.5	100
f_{PD} (GHz)	Frequency of PD	40	100
$N(-)$	$N \times N$ VMM scale	6	*up to* 1024
$T_{MZI}(-)$	Transmittance rate of MZI	0.9	0.99

of the state-of-the-art technology. Here, we discuss this from two perspectives: the impact of device size optimization and the power efficiency improvement by tolerating the calculation errors. Let us look at the evaluation results with the error rate of 0.3% on the lhs. Minimizing MZI components is a straightforward approach to optimization for device designers because it also reduces the latency of each MZI, maximizing f_{DPATH} as explained in Eq. (7). Unexpectedly, "device minimization ONLY" optimization (Case1) does not contribute to the power efficiency at all. The introduction of advanced parameters for f_{PS}, f_{PD}, f_{LS}, and T_{MZI} achieves 7.9–8.6× improvement, marked as Case2, but the impact is not sufficient. The important observation is that, as shown in Case3, the advanced value of the MZI size, N, significantly gains the power efficiency by a factor of 10.9–29.7×. This is because f_{VMM} is limited by f_{PS} rather than f_{DPATH} when the VMM size is small, and f_{DPATH} becomes critical if N is enlarged, i.e., the critical path that determines the maximum operation frequency of the VMM changes with N. The next discussion focuses on the accuracy of VMM calculation results. We can see an interesting result by comparing the power efficiencies of Case3 and Case6. By tolerating 62% VMM error rate on the rhs, a remarkable improvement of 178.1× is obtained. Note that we have confirmed that such an error rate does not cause any critical issues from the viewpoint of neural network inference applications. The above two scenarios suggest that we need to carefully consider the design parameters in a system-wide view and designers should co-optimize critical parameters that have significant impacts on power efficiency.

4. Challenges and future perspective

Although these ANN accelerators have significant potentials over electronic computing platforms, several issues remain to be solved. First, memory subsystems for caching are a critical challenge as the neural networks scale up to cope with real workloads, which generate a large amount of intermediate data for caching. One direction is to exploit electronic memories such as SRAM and DRAM for large-scale neural network executions and to integrate with the optical VMMs.[33] Second, establishing simulation methodologies and developing tools for cross-layer codesign are essential. Particularly, detailed power/performance modeling and fast design space exploration are required on nanophotonic accelerator designs. Third, considering optical-electrical heterogeneous computing models, revisiting computer architecture and defining hardware-software interfaces for nanophotonic ANN accelerations have to be considered.

C. Reservoir computing accelerator

Photonic reservoir computing has attracted growing interests as new artificial intelligence (AI) hardware. Reservoir computing is a class of recurrent neural networks whose connection weights among the network nodes are randomly fixed (referred to as "reservoir").[34,35] The training process of the connection weights is simplified in reservoir computing, and only the output weights are trained by learning. This approach has the advantages that the learning algorithm is simple and small computing power is required. The concept of reservoir computing is based on a mapping of an input signal into

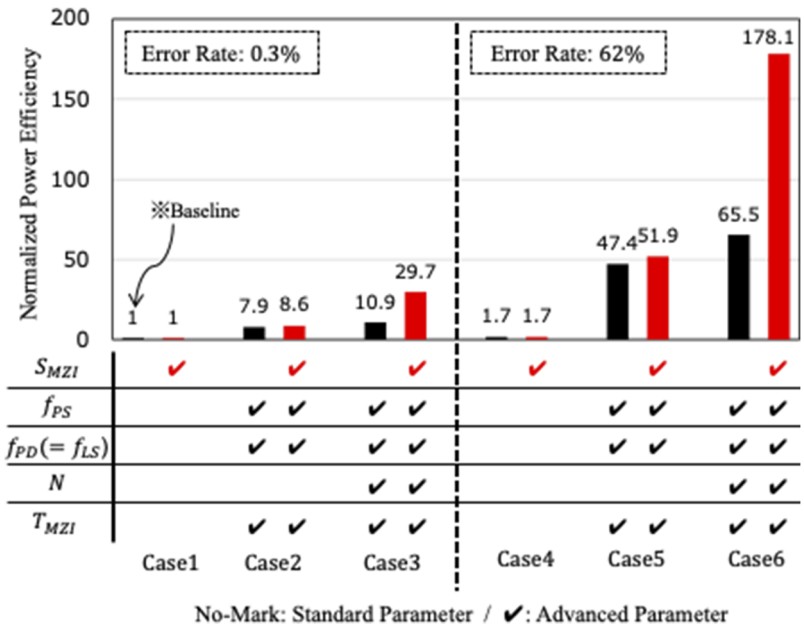

FIG. 9. Impact of design parameter optimization on VMM power efficiency.

a high dimensional space to facilitate classification and time-series prediction. There are two types of architecture: delay-based reservoir and spatial reservoir.

The concept of the delay-based reservoir computing using a single nonlinear system has been proposed.[36] This approach has simplified the hardware implementation of a virtual network based on a nonlinear dynamical system with time-delayed feedback. The outputs within the time-delayed feedback loop are sampled and treated as the virtual node states. The implementation of delay-based reservoir computing has been reported using optoelectronic systems,[37,38] passive nonlinear optical devices,[39] and laser dynamical systems.[40–43] Particularly, semiconductor lasers with time-delayed feedback are promising for high-speed implementation and high dimensional transformation of the input signal.[40] Recently, two-dimensional implementation of reservoir computing has been proposed using an SLM.[44]

The advantages of using photonic systems for reservoir computing include fast processing speed, parallel processing, and low-power consumption. The processing speed of photonic reservoir computing is determined by the transient response time of the reservoir, e.g., the relaxation oscillation time of semiconductor lasers at subnanoseconds. In addition, parallel processing can be enabled using the spatial and WDM techniques. Low-power consumption and small footprint can be achieved using photonic integrated circuits and Si photonics. Photonic reservoir computing can be advantageous over electronic reservoir computing in terms of speed and multiplexing capability for scaling.

Reservoir computing has been applied for time-series prediction and pattern recognition tasks.[36–44] Reservoir computing is considered to be suitable for the information processing that requires short-term memory of past input signals, the so-called fading memory. For example, the chaotic time-series prediction[45] and the tenth-order nonlinear autoregressive moving average (NARMA10)[36] are the well-known benchmark tests of the prediction tasks for reservoir computing. Spoken-digit recognition and nonlinear channel equalization are also commonly used as the benchmark tests of pattern recognition.[38,39]

1. Delay-based reservoir computing

Delay-based reservoir computing is structured with a nonlinear optical device in a time-delayed feedback loop as shown in Fig. 10. The time-delayed feedback loop with the delay time τ is considered as a reservoir, and a random temporal mask is required to introduce the weights between an input signal and the virtual nodes in the reservoir. The input signal is sent to the nonlinear device in the reservoir. In the reservoir, the temporal waveform in the delayed feedback loop is detected at the interval θ, and the amplitude of the temporal waveform is considered as virtual node states $x_i(n)$ for the nth input data. The output signal is calculated from the weighted sum of the virtual node states. The output weights W_i are trained in advance so that the output signal obtained from the reservoir can match the ideal output signal.

Figure 11 shows the implementation of reservoir computing using a semiconductor laser with optical feedback. The scheme is composed of three parts: the input layer, the reservoir, and the output layer. In the input layer, input data $u(n)$ are discretized and held at time T. A temporal mask is applied for each duration of T. The value of the mask is set at each interval θ, corresponding to the virtual-node interval in the reservoir. The feedback delay time τ in the reservoir is set to match the input holding time T, which is determined by the product of N virtual nodes and node interval θ (i.e., $T = N\theta$, also see Fig. 10).

Two semiconductor lasers, the drive and response lasers, are used. The drive laser is used to achieve consistent output from the response laser that is reproducible output with respect to the same input signal. The output of the drive laser is modulated by the input signal with the temporal mask, and the modulated signal is injected into the response laser. The drive injection signal to the response laser can suppress the DC noise of the response laser because the laser becomes more stable under the optical injection. The temporal waveforms of the response laser outputs are measured by a photodetector and a digital oscilloscope. The transient dynamics of the response laser is used as the reservoir output. The virtual nodes are determined from the transient response of the laser system for each interval θ within the feedback delay time τ. The reservoir is composed of virtual nodes $x_i(n)(i = 1, 2, \ldots, N)$ for the nth input data, and individual virtual nodes indicate different values to achieve high-dimensional space mapping. For postprocessing in the output layer, the output $y(n)$ for the nth input data is calculated as a linear combination of virtual nodes $x_i(n)$ with output weights, W_i, denoted as $y(n) = \sum_{i=1}^{N} W_i x_i(n)$, as defined in Sec. III B 1. The output weights W_i are optimized by minimizing the mean-square error between the target function and the reservoir output $y(n)$ in advance. The linear least-squares method is used for learning W_i with the training data.

The chaotic time-series prediction task[45] is used to evaluate the performance of the reservoir computing. The aim of this task

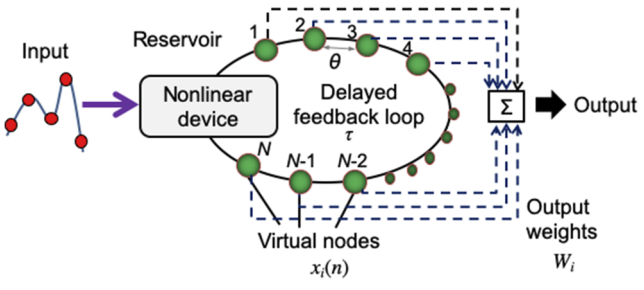

FIG. 10. Scheme of delay-based reservoir computing.

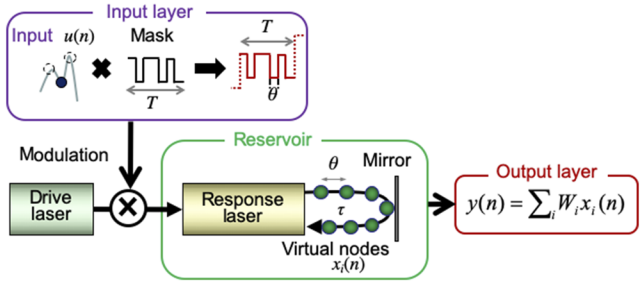

FIG. 11. Scheme of delay-based reservoir computing using a semiconductor laser with time-delayed feedback. Adapted from Ref. 42.

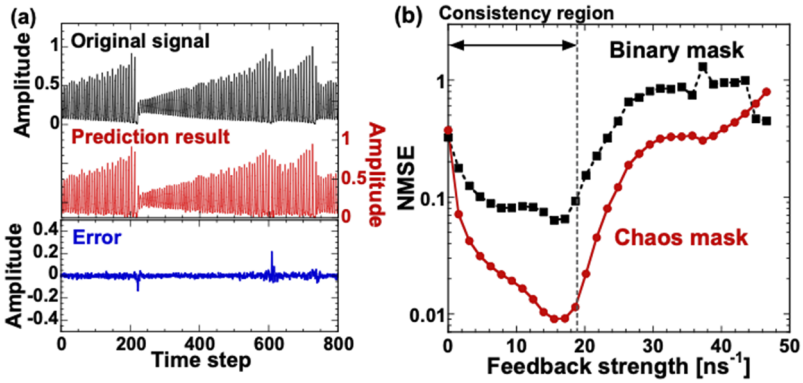

FIG. 12. (a) Result of chaotic time-series prediction task. The original input signal, prediction result by reservoir computing, and prediction error between them are shown for the *n*th input data. (b) Normalized mean-square errors (NMSEs) as a function of the feedback strength of the response laser for the binary (the black dotted curve with squares) and chaos (the red solid curve with circles) mask signals. The consistency region is shown by the arrow. Both figures are adapted from Ref. 42.

is to perform single-point prediction of chaotic data, which are generated from a far-infrared laser. Figure 12(a) shows the temporal waveforms of the original input signal, the prediction result obtained from the reservoir computing, and the error between them for the *n*th input data. It is found that the prediction result is very similar to the original input signal, and a small error is obtained. A normalized mean-square error (NMSE) of 0.008 is obtained in Fig. 12(b). The chaotic time series prediction is successful using the reservoir computing scheme. Other tasks such as NARMA 10 and the nonlinear channel equalization have also been succeeded in this scheme.

The temporal mask signal plays an important role as the input layer-reservoir connections for reservoir computing in Fig. 11. A binary random mask signal is used in many cases, consisting of a piecewise constant function for each interval θ, with a randomly modulated binary sequence $\{-1, 1\}$ with equal probabilities. On the contrary, a chaos mask signal can be used, generated from the response laser with optical feedback, but without drive injection. Figure 12(b) shows the NMSE when the feedback strength of the response laser is changed for the chaos and binary mask signals. Smaller NMSEs are obtained for the chaos mask signal, and the NMSE is minimized at the boundary of the consistency region. It is found that the performance of reservoir computing can be improved using the chaos mask signal because the response laser generates complex transient dynamics, and a variety of node states can be obtained for the chaos mask signal.

The consistency property,[46] which is the ability to produce the same output for repetitive input injection, is required for reservoir computing. The feedback strength in the delayed feedback loop is set to be weak to avoid chaotic outputs because the consistency property is satisfied under the steady state of the laser output without drive injection. In fact, better performance of reservoir computing has been achieved in the boundary between the steady state and the chaotic state [known as "the edge of chaos" as shown in the consistency region indicated by the arrow in Fig. 12(b)].

Different implementations using nonlinear optical devices have been reported, such as optoelectronic systems,[37,38] passive nonlinear devices,[39] and a semiconductor laser with optical feedback.[40–43] The complexity of the output of the reservoir depends on the nonlinearity of the optical devices. For example, the saturation of the input power is one of the simplest nonlinear functions when using a semiconductor optical amplifier as the reservoir. The sinusoidal function is implemented for electro-optic modulators. More complex dynamics can be obtained for the semiconductor laser with optical feedback, which is governed by the rate equations known as the Lang-Kobayashi equations.[47,48]

2. Spatial reservoir computing

Spatial implementation of reservoir computing has been proposed.[44] In this scheme, the reservoir consists of many nonlinear nodes in a spatially distributed network. Figure 13 shows the experimental setup of the spatial implementation of reservoir computing. The optical output reflected from each pixel (or a group of the pixels) on the SLM is considered as a node state of the reservoir. The output of a semiconductor laser is sent to the SLM, and the phase of the injection light is modulated at each pixel. The reflected light from the SLM is sent to a diffractive optical element (DOE) and an external mirror to realize the connection among the nodes. The reflected light from the external mirror is sent to a digital camera, and the detected signal is used to modulate the pixels on the SLM. In addition, the light from the SLM is sent to a digital micromirror device (DMD) to

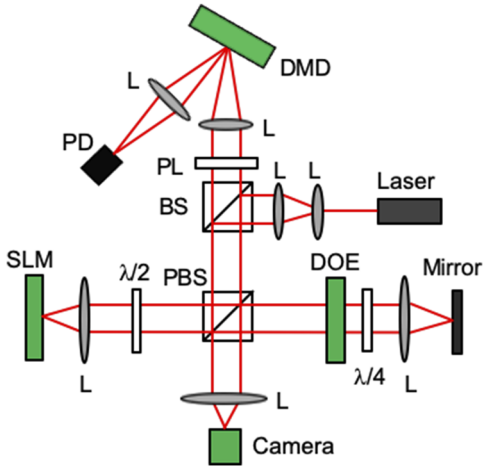

FIG. 13. Experimental setup of spatial reservoir computing. BS, beam splitter; DOE, diffractive optical element; DMD, digital micromirror device; SLM, spatial light modulator; L, lens; λ/2, half waveplate; λ/4, quarter waveplate; PBS, polarization beam splitter; PD, photodetector; and PL, polarizer. Adapted from Ref. 44.

implement the output weights of the reservoir. The weighted sum of the node states is optically calculated by focusing the optical output from the DMD. 900 nodes have been implemented in this experiment, and chaotic time-series prediction has been demonstrated successfully with the NMSE of 0.013.[44]

The pros and cons of delay-based and spatial schemes are compared as follows. For the delay-based scheme, only a single nonlinear device with time-delayed feedback is required, and the implementation can be simplified for a large number of the virtual nodes. However, the complex preprocessing and postprocessing are required to add the temporal mask to the input signal and to calculate the weighted sum of the virtual node states. For the spatial scheme, by contrast, no complex preprocessing and postprocessing are required, and the on-line implementation can be achieved. However, a large number of nodes are required for the network, and it is difficult to achieve the hardware implementation with too many nodes under stable operation. In addition, delay-based reservoir computing might require high-power. On the other hand, spatial implementation of reservoir computing using photonic integrated circuits can be useful for low-power consumption. Recently, a hybrid technique of the delay-based and spatial schemes has also been reported.[49]

For the spatial scheme, coupled laser arrays[50] and quantum-dot micropillar lasers[51] are very promising devices for reservoir computing, and each laser can be considered as a node of the network. High-speed and low-power consumption operation of reservoir computing is expected using these photonic devices.

3. Challenges and future perspective

From the engineering point of view, it is necessary to develop fast, compact, and low-power consumption reservoir computing using photonic devices. For the speedup, real-time implementation of reservoir computing has been demonstrated using a FPGA,[52] and chaotic time-series prediction has been achieved on real-time basis. In addition, parallel processing using WDM has been proposed for reservoir computing.[53] For compact implementation, a photonic integrated circuit (PIC) has been used for reservoir computing. Figure 14 shows the configuration of the PIC used for delay-based reservoir computing.[54] The PIC consists of a semiconductor laser, a semiconductor optical amplifier, a phase modulator, and an external mirror whose external cavity length is ~10 mm. The number of virtual nodes in the reservoir is limited due to the short external cavity length in this scheme. Therefore, the temporal waveform corresponding to multiple delay times is used to increase the number of virtual nodes. The performance of the chaotic time-series prediction task has been successfully demonstrated using this scheme with the prediction error (NMSE) of 0.109.[54]

For low-power consumption, a Si-based PIC has been also introduced to implement spatial reservoir computing.[55] The PIC

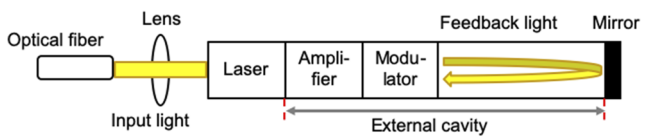

FIG. 14. Photonic integrated circuit (PIC)-based delay-based reservoir computing. Adapted from Ref. 54.

consists of splitters, combiners, and optical waveguides, and 16 nodes are implemented in the PIC within the footprint of 16 mm^2. All the optical components are passive and linear, and the nonlinearity is introduced externally as the saturation of the detection parts. Fast implementation of 5-bit header recognition has been demonstrated at 12.5 Gbit/s.[55]

Reservoir computing is suitable for time-dependent information processing, such as time-series prediction and speech recognition. Recently, other applications have been reported, including bit detection based on pattern recognition in a noisy optical communication system[56,57] and radar signal prediction.[39] In Refs. 56 and 57, the feasibility of postprocessing optical PAM4 signals has been demonstrated for optical communications using photonic reservoir computing schemes. Recently, multilayer (deep) reservoir computing schemes using multiple reservoirs have been demonstrated to improve the performance of reservoir computing.[58–60] The standard benchmark test for image recognition, known as MNIST, has been solved with high correct recognition rate using deep reservoir computing.[60] In addition, "deep" reservoir computing has been proposed and the standard benchmark test for image recognition, e.g., MNIST has been solved with a high correct recognition rate.[60] In addition, the combination of another functionality with computing in photonic devices would be one of the next challenges. For example, reservoir computing could be combined with optical sensors, and compact sensors with self-data analyzer using reservoir computing could be implemented.[61,62] Further investigation of reservoir computing is still required, and it will open a new research direction and engineering applications of photonic accelerator.

D. Nanophotonics-based pass-gate logic accelerator

It is well recognized that the progress of CMOS transistors in terms of the latency has been already leveled off. This issue is attributed to the fact that RC delay of a CMOS transistor with minimum wiring is saturated around 10 ps or so, irrespective to the reduction of the gate size.[63–65] We discuss the possibility for breaking this barrier by introducing optics into a processor because optical circuits are not limited by the RC delay.[66,67] This is particularly interesting in terms of the pass gate logic, where the computation time is determined by the signal propagation. A challenge is not to intend for all-optical computing, but it is rather optoelectronic computing where optics is implemented within electric circuits. Figure 15 illustrates a generic optoelectronic computation circuit. It is assumed that the input data E_{in} are fed into the optoelectronic circuit as an electric signal, and the computation is done as a result of light propagation via this optoelectronic circuit, which we call optical pass gate (OPG) circuit. At the last stage, the light output is evaluated by the electric circuits via OE conversion, and the final result is generated as an electric signal E_{out}. The latency of this type of circuits may become extremely short if the light propagation length is short enough, and the overhead of OE/EO conversions is sufficiently small. When the optical component size is in a centimeter scale, this is typical at the old era of optical computing research in the 1980s or 1990s. Hence, the path delay is around 100 ps in Si, which is much longer than the CMOS delay. Therefore, the size reduction of optical circuits is crucial to enjoy the merit of this computation scheme.

Nanophotonics will play a key role for reducing the footprint of circuits, together with the energy reduction.[68] The size of the

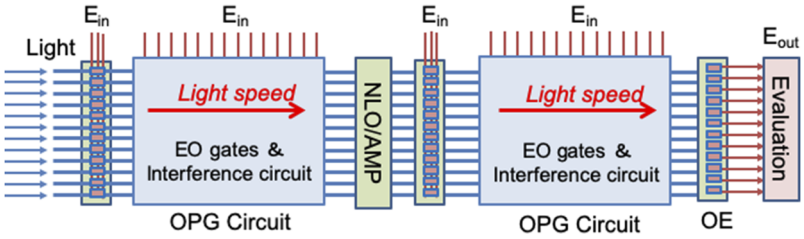

FIG. 15. A generic optoelectronic computation circuit. The computation should be done by light propagation from left to right and finally evaluated by an electronic circuit via OE conversion.

latest nanophotonic gates is about tens of micrometer or shorter (a few micrometer in extreme cases[69]), indicating that the path delay is subpicosecond, which is much shorter than the CMOS delay. Of course, the switching delay of optoelectronic devices is at most comparable to that of the CMOS; therefore, the total computation time of the circuit, as in Fig. 15, will be dominated by the path delay that is the signal propagation time from west to east. In addition, the OE/EO conversions in Fig. 15 play an essential role in this type of optoelectronic computing. Conventionally, OE/EO conversions are considered the most inefficient and energy-hungry ingredients in optical information processing. There has been a kind of consensus that it is better to reduce or eliminate the OE/EO conversion nodes if we want to enjoy the merit of optical processing. However, the recent progress of OE/EO converters by use of nanophotonics could change this paradigm, and the optoelectronic computation scheme becomes highly promising as is discussed in Sec. III D 1.

1. Recent progress of OE/EO conversion

We will introduce a recent breakthrough in terms of the reduction of the capacitance for OE/EO converters by focusing on photonic crystal (PhC) PD and PhC EO modulator (EOM). Recent progress of nanophotonics has enabled astonishing reduction in terms of the consumption energy together with the delay. The capacitance of conventional optoelectronic devices such as PDs and EOMs has been around hundreds of femtofarad to picofarad range, which is much greater than a femtofarad level of electronic devices such as CMOS transistors. This large capacitance is primarily due to the poor light confinement ability in conventional optoelectronic device technologies. Meanwhile, the recent nanophotonics progress has achieved a dramatic reduction of the capacitance by orders of magnitudes as shown in Fig. 16. Especially, a capacitance of a femtofarad

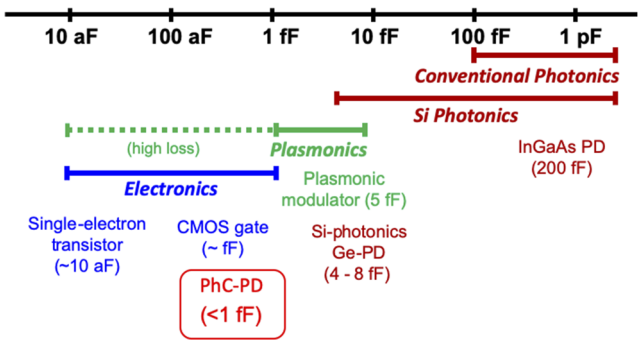

FIG. 16. Comparison of the capacitance for various devices, including electronic,[63,70] photonic,[71–73] and plasmonic devices.[74]

level for PDs and EOMs has been achieved by employing PhCs.[75] A PhC is an artificial dielectric structure whose refractive index is periodically modulated in a 100-nm scale.[76,77] PhCs can be fabricated by semiconductor nanofabrication techniques such as CMOS fabrication process, and it has been well established that PhCs enables unprecedented strong light confinement.[77]

A conventional OE converter consists of a combination of a PD and a *trans*-impedance amplifier (TIA).[66] The latter requires generation of sufficient voltage to drive subsequent CMOS transistors. The energy consumption of TIA is typically several hundred fJ/bit, which is too large for our application. However, if ultrasmall capacitance PDs (such as in a femtofarad level) are available, we can replace an energy consuming TIA with a passive load resistor (1–10 kΩ). Sufficiently large voltage can be generated by a photocurrent directly from the resistor without sacrificing the operation speed. Recently, PhC PDs with a capacitance of 0.6 fF have been demonstrated, which are already comparable to that of CMOS.[68,78] As shown in Fig. 17, these PhC PDs retain high performance mostly comparable to conventional high-speed PDs with much bigger dimension. By monolithically integrating this PhC PD with a load resistor of 8.8 kΩ, very high light-to-voltage conversion efficiency of 4 kV/W has been demonstrated, which is even greater than that for commercially available PD-TIA photoreceivers.

Very recently, the PhC PD has demonstrated to be operated at a *forward-biased* condition even at 20 Gbit/s.[78,79] This result suggests that when this PD is integrated with a substantially large resistor, it can operate at a zero-bias condition. In this condition, we would be able to remove even the bias power supply from OE converters. Although this is still a preliminary result, it indicates a very promising direction, that is, the OE conversion can be done without any electric power. The small capacitance is vital even more substantially for the EO conversion.

The energy consumption of EOM is dominated by the charging energy $\propto CV^2$ with the capacitance C and applied voltage V; therefore, the small capacitance is crucial. A high-speed operation at 40 Gbit/s of PhC waveguide-type EOMs in Fig. 18, whose capacitance is as small as 13 fF has been reported. The energy consumption is smaller than 2 fJ/bit, which is the lowest value for waveguide-type EOMs.[80] This value could be further reduced by introducing a nanocavity-based design because the capacitance of the nanocavity-based devices could be similar to the capacitance of the PhC-PDs.[81] From these achievements, it becomes possible to greatly reduce the energy consumption of OE/EO conversions by employing PhC technologies. Consequently, armed with nanophotonic OE/EO conversions, we may have to reconsider the border between electronics and photonics in information processing. This suggests that OE/EO conversions can potentially be employed in more and more fine-grained levels, which support the computation scheme shown in Fig. 15.

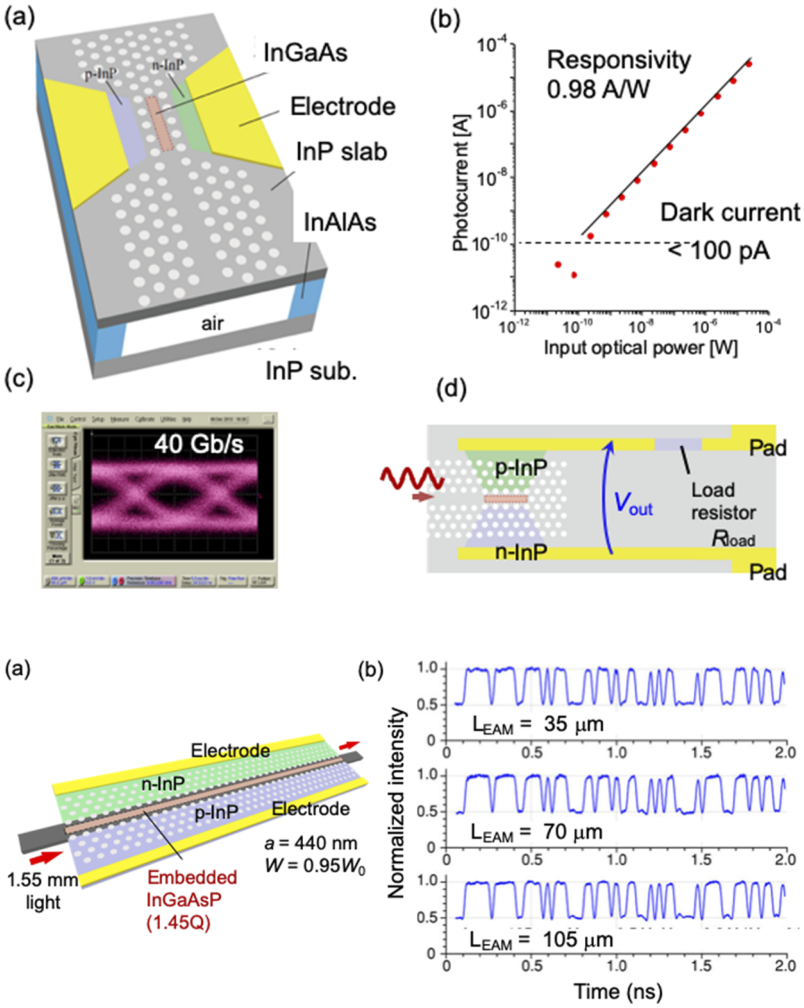

FIG. 17. (a) A schematic of a PhC PD. The gray part is an InP photonic crystal waveguide. The red part is an InGaAs absorber which is located within a lateral p-i-n junction. The length of the absorber region is 1.7 μm, and the device capacitance is 0.6 fF, estimated by the three-dimensional electromagnetic simulation. (b) Measured photocurrent vs the input optical power, showing a high responsivity of 0.98 A/W and a substantially small dark current <100 pA. (c) Measured eye diagram at 40 Gbit/s NRZ signal input. (d) A schematic of a PhC PD integrated with a load resistor. All the figures are adapted from Ref. 75.

FIG. 18. (a) A schematic of a waveguide-type PhC EOM and (b) modulation characteristics at a bit rate of 40 Gbit/s of PhC EOMs with L = 35 μm, 70 μm, and 105 μm. Adapted from Nozaki *et al.*, APL Photonics **3**, 046101 (2018). Copyright K. Nozaki *et al.*, APL Photonics **3**, 046101 (2018). Copyright 2018 Author(s), licensed under a Creative Commons Attribution 4.0 License.[79]

2. Optical pass-gate logic

We introduce a specific example of the optoelectronic computation scheme in Fig. 15. Figure 19(a) shows a logic gate, the so-called pass-gate logic, in comparison with a conventional CMOS logic of AOI (AND-OR-Inverter logic) in Fig. 19(b). With the conventional AOI logic, the computation is done by a sequential operation of CMOS gates. The total computation time is a simple sum of $N\tau_{switch}$ + $N\tau_{path}$, where N, τ_{switch}, and τ_{path} are the number of gates, the switching delay per gate, and the path delay per gate, respectively. By contrast, with the scheme shown in Fig. 19(b), all possible answers are injected from the lhs, and the input data select the path. Only the correct answer arrives at the exit. It is noteworthy that with this scheme, all the gates can be switched simultaneously, and thus, the total computation time is a sum of τ_{switch} + $N\tau_{path}$. Since the switching delay is not accumulated, the computation time will be dominated by the path delay when N becomes large. In fact, this scheme is essentially a circuit representation of the binary decision diagram (BDD), which is similar to the pass-transistor logic or the one employed in FPGA. It is well known that BDD can represent any AOI logic, but this scheme has rarely been used as low-latency circuits because a sequential connection of CMOS, which consists of the critical path of the calculation, leading to a substantial RC delay which is scaled as the square of the number of gates N^2. However, if we realize this circuit with optical gates, such as EO-MZI switches, the path delay is simply the time required for the light propagation,

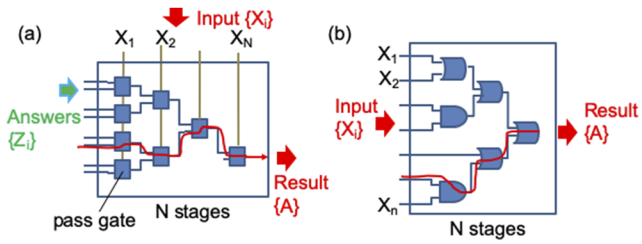

FIG. 19. (a) Optical pass-gate logic and (b) AND-OR-Inverter logic conventionally employed in CMOS computing.

scaled as N. This immediately means that if we embody this circuit with nanophotonic gates whose lengths are shorter than 100 μm, the path delay per gate could be as short as 1 ps.

Recently, this BDD scheme has attracted much attention from optic communities. Hardy *et al.* proposed to employ 2 × 2 EO-MZI switches to construct various logic circuits, which they call optical directed logic, especially promising for reversible computing.[82] A specific example of application has been proposed to employ the BDD to optical adders and ripple-carry adder.[83–85] A conventional parallel full adder is realized by a set of AND, XOR, and OR gates, and it is possible to transpose such circuit into BDD.

3. Challenges and future perspective

One of the immediate applications of the light-speed computation is an optical adder by the BDD-like circuits. Since the critical path of the ripple carry adder can be "optical," it is relatively easy to implement the scheme shown in Fig. 15 if it is in a small scale. The technical challenge lies in the fact that the BDD scheme generally produces relatively large and complicated circuits, and thus, it is important to optimize the circuit design to reduce the number of gates. Recently, a simplified design of an optical adder was proposed, which requires only a single MZI switch per digit,[85] in which the functionalities of EO-MZI switches is fully utilized, and employs the degree of freedom for the wavelength to express the logical negation. This shows that although the BDD-type circuit formalism has been known for decades by circuit architects, we will have to expand the concept of the BDD and invent novel circuit designs appropriate for the optical pass-gate logic.

The full adder is only an example, and the light-speed computation scheme can be extended to various computation circuits. Apparently, the pattern matching is very suited for this scheme. Another important direction is to pursue optical interference circuits. Recently, simple multiport interferometers were proposed to enable Boolean logic functionalities.[86] In addition, an interference matrix circuit made of EO-MZI switches can perform VMM. The VMM is widely used in various applications such as the ANN in Sec. III B 1.

E. Decision making accelerator

Decision making is to perform adequate judgments in uncertain environments which are widely applied to information and communications technology, including dynamic and efficient resource assignments in network infrastructures,[87,88] search functions,[89] and the foundation of reinforcement learning.[90] A fundamental and important issue in decision making is formulated by the multiarmed bandit (MAB) problem where the problem is to maximize the player's rewards from many slot machines whose reward probabilities are unknown.[90] To find the best machine, sufficient exploration is necessary; however, too much exploration may accompany significant loss. Moreover, the best machine may change over time. On the other hand, a too quick decision may miss the best machine. Such a difficult trade-off has been known as *exploration-exploitation dilemma*. In this section, decision making refers to solve MAB problems.

Recently, MAB problems have been implemented with photonic technologies unlike conventional computer algorithms performed in digital computers[91–93] to pave the way for breaking the limitations of conventional approaches such von Neumann bottleneck,[94] energy efficiency, and operation speeds. In particular, by pursuing the ultimate performances of photonic technologies, one can design novel system architectures and functionalities. In Sec. III E 1, single-photon decision maker is described by utilizing wave-particle duality of light quanta. The principle is transformed to laser systems where chaotically oscillating ultrafast dynamics are utilized, exploiting the superior scalability by time-domain multiplexing as described in Sec. III E 2.

It would be noteworthy that photonic decision making shows a clear departure from recent photonic computing such as photonic Ising machines,[95] optical reservoir computing[37,40] in Sec. III C, and the ANN[44,96] in Sec. III B in terms of its intended functionalities and architectures. The goal of the above-mentioned systems is solving combinatorial optimization or recognition/classification tasks involving a certain amount of training data sets, which does not apply for the photonic decision maker. However, the photonic decision makers and photonic solution searchers are in a complemental relation. Indeed, for example, Kanno *et al.* demonstrated a composite system of a laser-chaos decision maker and photonic reservoir computing; the photonic decision maker dynamically switches the reservoir to be used for prediction. That is, the decision maker adaptively chooses an optimized reservoir among pretrained multiple ones depending on the incoming signal train.[97]

1. Single-photon-based decision making

The MAB becomes extremely difficult to solve when the number of potential candidates of decisions or the number of slot machines increases. Even the two-armed bandit problem, which involves only two slot machines, Machine 0 and Machine 1, is difficult to solve. Intuitively, one may readily conduct decision making when the selected machine mostly wins over recent trials. However, in reality, one can be easily fooled by such incidental events because the other machine may be the better machine.

The single-photon decision maker directly benefits from the wave-particle duality of light quanta for the exploration actions.[98,99] As shown in Fig. 20(a), a single photon emitted from a nitrogen-vacancy or NV-center in a nondiamond is utilized. A linearly polarized single photon impinges on a polarization beam splitter (PBS) after passing through a polarizer. If the photon via the vertical direction is detected by an avalanche photodiode (APD_0), the decision is immediately to select Machine 0. When the photon is detected by APD_1 via the horizontal polarization, the decision is to choose Machine 1. When a linearly polarization single photon polarized 45° with respect to the horizontal impinges on the PBS, the probability of photon detection by APD_0 or APD_1 is 50:50, meaning that the decision making is completely an exploration action. When the polarization of the input single photon is almost vertical, the photon is highly likely observed by APD_0 whereas *almost* horizontally polarized single photon will be detected by APD_1. Therefore, the strategy of decision making is to control the single photon polarization toward the better slot machine, which is experimentally realized by the halfwave plate ($\lambda/2$) before photons go through the PBS. It should be noted is that the function of "checking-the-other-machine" is physically realized by the probabilistic attribute of single photons; a nearly horizontal single photon is mostly detected by APD_1, but it is sometimes detected by APD_0. Such a property cannot be achieved if the input photon is "classical" when the physical quantity

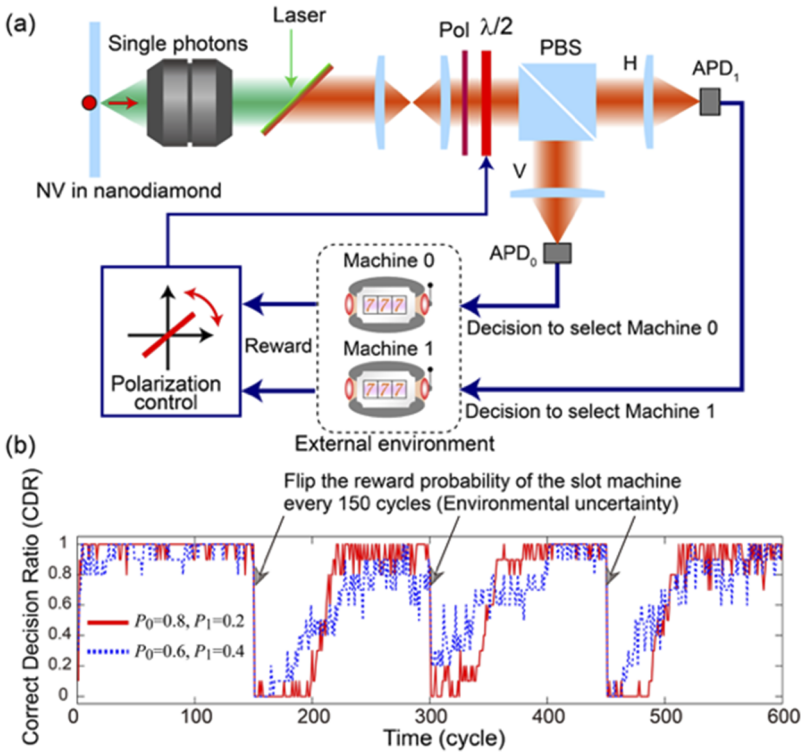

FIG. 20. (a) Architecture of single photon decision maker for solving two-armed bandit problem (b) Experimental demonstrations of decision-making adapting to uncertain environment. Adapted from Naruse *et al.*, Sci. Rep. **5**, 13253 (2015). Copyright 2015 Author(s), licensed under a Creative Commons Attribution 4.0 License.

measured by photodetectors are the ratio of light intensity between APD_0 and APD_1, and hence, one additional step is necessary to determine the decision by using random numbers generated in the computer.[98]

In the experiment, the arrival timing of photons at APDs was characterized by a time-correlated single-photon detection system.[100] The two slot machines were implemented in a computer using pseudorandom numbers. Based on the betting results, the angle of the halfwave plate was mechanically controlled. In Fig. 20(b), the horizontal axis indicates the number of cycles of the slot machine play, while the vertical axis depicts the correct decision ratio (CDR) which is the ratio of selecting the higher reward probability slot machine over total 10 repetitions. In the first 150 cycles, the reward probabilities of Machines 0 and 1 were given by $P_0 = 0.8$ and $P_1 = 0.2$, respectively. Therefore, choosing Machine 0 is the correct decision. The CDR depicted by the solid curve in red quickly approaches to unity, meaning that correct decision-making has been made. In every 150 cycles, the reward probabilities of machines are intentionally inverted to implement environmental uncertainty. That is, P_0 becomes 0.2, and P_1 is changed to 0.2. Accordingly, the CDR drops at the 151st cycle, but it gradually recovers to high values, meaning that the autonomous adaptation to environmental changes is realized. The dotted line in blue shows the results when the reward probabilities are given by 0.6 and 0.4, which is a tougher decision problem since the difference between the reward probability is smaller (0.6 − 0.2 = 0.4) than the previous case (0.8 − 0.2 = 0.6). Correspondingly, the CDR recovery is slower and shows relatively lower values, compared with the previous case in red. Nevertheless, we can observe clearly the autonomous decision making.

By pursuing the subwavelength-scale ultrasmall attributes of optical near-fields,[101] nanophotonic decision making has been also investigated theoretically in Ref. 102 and experimentally in Ref. 103. The energy transfer via the optical near-field is configured by control light so that the energy is selectively transferred to either of the quantum dots corresponding to decisions 0 or 1.

2. Laser-chaos-based ultrafast decision making

The generation of chaos in semiconductor lasers has been extensively studied.[104,105] Lasers become unstable when a fraction of the output light is fed back to their cavity, leading to chaos, for example. Ultrafast physical random generations have been experimentally demonstrated[106,107] by utilizing the high bandwidth attributes of chaotic lasers. Furthermore, the nonlinear dynamics of chaotic lasers, including its ultrafast transient process, are a fascinating resource for reservoir computing.[40] The application of laser chaos to decision making is one of the most recent topics in chaotic lasers.

As schematically shown in Fig. 21(a), a light intensity level is sampled from the laser chaos signal train, followed by thresholding denoted by *TH*. When the signal level is larger than *TH*, the decision is immediately made to select Machine 0, whereas when the signal is lower than *TH*, the decision is to select Machine 1. The threshold level is reconfigured based on the betting result. Assume, for instance, that the threshold is high enough; the sampled signals will mostly be smaller than the threshold; hence, the decision is mostly Machine 1. However, chaotically oscillating input light could sometime be even larger than the threshold; that is, the decision to select Machine 0 could occasionally be made, providing the ability of "checking-the-other-machine."

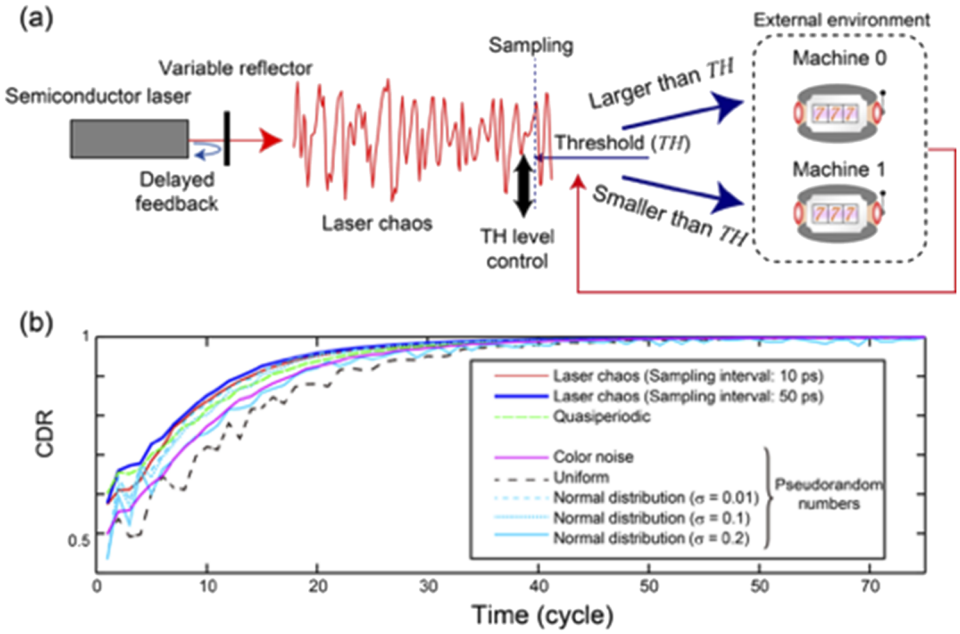

FIG. 21. (a) Architecture of decision maker based on laser chaos. (b) Laser chaos sampled with 50 ps interval provides the fastest adaptation. Both figures are adapted from Naruse *et al.*, Sci. Rep. **7**, 8772 (2017). Copyright 2017 Author(s), licensed under a Creative Commons Attribution 4.0 License.

The curve in blue in Fig. 21(b) shows the evolution of CDR when the chaotic signal is sampled with 50 ps interval that shows the promptest adaptation to unknown environments. After about 20 cycles, the CDR becomes larger than 0.9, meaning that that the latency from the initial status to the correct decision is about 1 ns. The sampling interval 50 ps that provides the best decision-making performance exactly coincides with the negative autocorrelation inherent in the chaotic time series.[108] At the same time, although a quasiperiodic signal contains larger negative maximum in the autocorrelation, the performance of decision making is not good. Assuming that pseudorandom numbers and color noise were available in such a high-speed domain, which is extremely difficult, the laser chaos could outperform these alternatives as shown in Fig. 21(b). Here, the color noise containing negative autocorrelation was calculated based on the Ornstein–Uhlenbeck process using white Gaussian noise and a low-pass filter[109] with a cut-off frequency of 10 GHz. Such an observation implies that the temporal structure of irregular signals in laser chaos is positively affecting the performance of decision making, which is further discussed in the following.

Not just two-armed bandit problems, next we consider N-armed bandit problems where $N = 2^M$ with M. Let the N slot machines be distinguished by natural numbers ranging from 0 to $N - 1$ represented in an M-bit binary code given by $S_1S_2\cdots S_M$ with S_i ($i = 1, \ldots, M$) being 0 or 1. For example, when $N = 8$ (*or* $M = 3$), the slot machines are numbered by $S_1S_2S_3 = \{000, 001, 010, \ldots, 111\}$ in Fig. 22(a). The reward probability of Machine i is represented by P_i ($i = 0, \ldots, N - 1$). This scalable decision making by chaotic lasers have been demonstrated in Ref. 97. The decision is determined bit by bit from the most significant bit (MSB) to the least significant bit in a pipelined manner. For each of the bits, the decision is made based on a comparison between the measured chaotic signal level and the designated threshold value. First, the chaotic signal $s(t_1)$ measured at $t = t_1$ is compared to a threshold value denoted as TH_1 in Fig. 22(b). The output of the comparison is immediately the MSB of the decision. When $s(t_1)$ is less than or equal to the threshold TH_1, the decision is that the MSB of the decision is 0. Otherwise, the MSB is determined to be 1. The chaotic signal $s(t_2)$ measured at $t = t_2$ is subjected to another threshold TH_{20}. If $s(t_2)$ is less than or equal to the threshold TH_{20}, the second MSB of the decision is 0. Otherwise, the second-most significant bit of the decision is set as 1. All of the bits are determined in this manner.

Four kinds of chaotic signal trains were generated, referred to as Chaos 1, Chaos 2, Chaos 3, and Chaos 4 by varying the reflection by the variable reflector by letting 210, 120, 80, and 45 μW of optical power be fed back to the laser, respectively. A quasiperiodic signal train was also generated by the variable reflector by providing a feedback optical power of 15 μW. In addition, computer-generated, uniform pseudorandom numbers and color noise were examined for comparisons. We applied the decision-making strategy to bandit problems with two, four, eight, 16, 32, and 64 arms. We assigned the reward probabilities to the N machines so that the difficulty maintains coherence. The detailed settings are described in Ref. 110. Figure 22(b) summarizes the results of the 16-, 32-, and 64-armed bandit problems, respectively. The curves in red, green, blue, and cyan show the CDR evolution obtained using Chaos 1, 2, 3, and 4, respectively, while the curves in magenta, black, and yellow depict the evolution obtained using quasiperiodic signals, pseudorandom numbers, and color noise, respectively. From Fig. 22(b), Chaos 3 provides the promptest adaptation to the unity value of the CDR, whereas the nonchaotic signals (quasiperiodic, pseudorandom, and color noise) yield substantially deteriorated performances. The number of cycles necessary to reach a CDR of 0.95 increases as the number of bandits in the form of the power-law relation aN^b, where a and b are approximately 52 and 1.16, respectively, indicating that the successful operation of the proposed scalable decision-making principle using laser-generated chaotic time series.

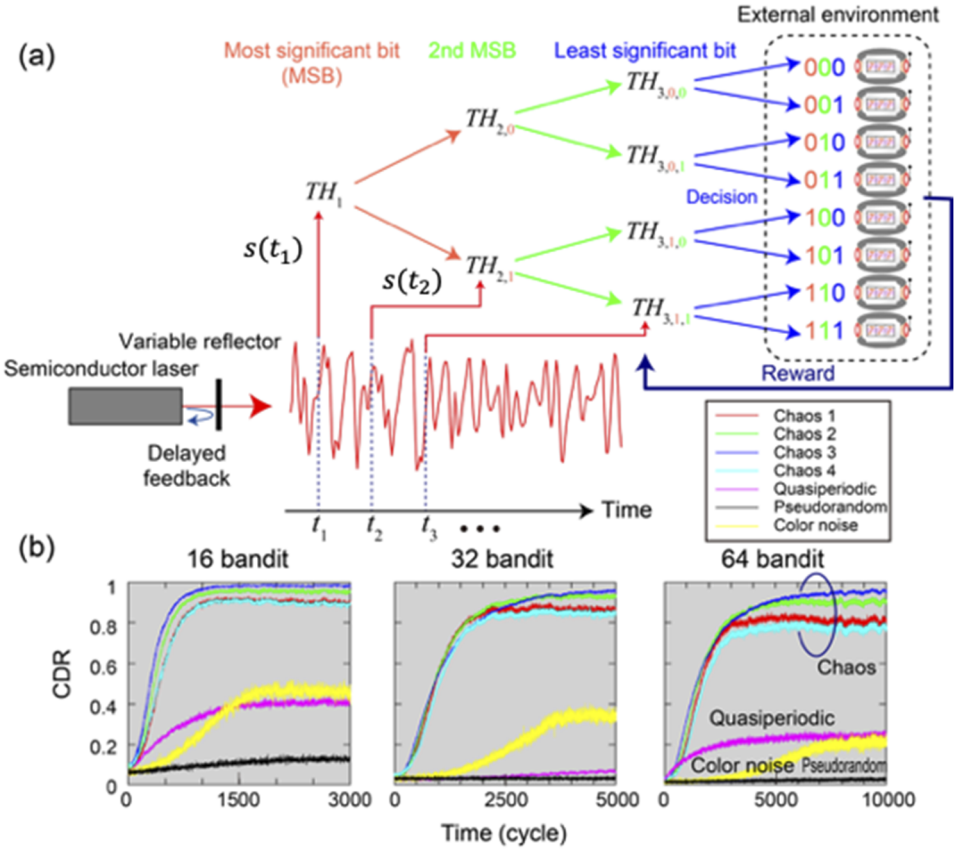

FIG. 22. (a) Scalable decision making by pipelined thresholding of consecutively sampled laser chaos sequence. (b) Demonstration of solving 16-, 32-, and 64-armed bandit problems where chaos achieved superior performances than other random sequences. (c) Diffusivity analysis of the irregularity inherent in the time series. Chaotic sequences yield wider exploration in the phase space. All the figures are adapted from Naruse *et al.*, Sci. Rep. **8**, 10890 (2018). Copyright 2018 Author(s), licensed under a Creative Commons Attribution 4.0 License.

3. Challenges and future perspective

Photonic decision making is an emerging research area; hence, there are vast and versatile possibilities from a variety of perspectives. Regarding physical principles, we can further utilize the rich physics of photons for decision making and not just those reviewed above. For instance, Mihana *et al.* demonstrated the usefulness of the leader-laggard laser synchronization between multiple lasers for decision making where the impact of networking and synchronization in network of lasers is exploited.[111] From the viewpoint of photonic devices, recent advancements in PIC indicate on-chip photonic decision makers. In fact, Homma *et al.* demonstrated a photonic decision maker using a microring laser where the spontaneous and chaotic oscillatory dynamics between clockwise and counterclockwise light propagations are utilized for making decisions.[112] Regarding nanophotonic technologies, Nakagomi *et al.* demonstrated in writing nanoscale patterns on the surface of photochromic single crystal via optical near-field, meaning that the function of memorizing past events, an important aspect for making decisions, has been shown.[113] Also, Mihana *et al.* examined theoretically the impact of the memorization of past history for decision making in dealing with uncertain environments.[114]

A decision making system involves external environments; hence, the interface between photonic decision maker and external systems is an important issue. Conventionally, when we deal with ultrafast optical signals, for instance, in optical communications and computing applications,[95] multiplexing/demultiplexing and Application Specific Integrated Circuit (ASIC)/FPGA circuits are typically employed to resolve the mismatch of the fast optical signals and slow electrical signal processing units. However, such an architecture implies that the potential abilities of photonics could be severely limited by conventional electronics. Therefore, integration of photonic and electronic devices is another important subject. Indeed, recent Si CMOS device technologies realized 300 GHz operating speed for THz transmitters and receivers,[115] implying that a certain class of analog signal processing is achievable in the high-speed regime including photonic decision makers. The scalability of photonic decision maker has been demonstrated up to 64 bandits by a time-domain multiplexing strategy as shown in Sec. III E 2.[110] Note, however, that there exist other interesting possibilities for scaling, including the utilization of spatial parallelism on the basis of above-mentioned PIC[112] and nanophotonic technologies,[113] WDM, mode-division multiplexing, and combinations of these principles.

Establishing theoretical backgrounds for photonic decision making is also an important subject. Conventional principles are based on digital computing algorithms which are supported by fundamental computing studies by Turing and von Neumann. On the other hand, the systems under study herein are coupled with natural phenomena, such as a single photon or laser chaos, accompanying irregular signals with a view to accomplish performance enhancement or acceleration as a whole. For such a purpose, a novel theoretical framework is necessary. A category theoretical analysis

has been demonstrated for clear understanding of complex interdependencies between multiple subsystems in photonic decision makers.[116] Such a categorical approach is also applied to an analysis of choice-based learning in brain.[117] Saigo *et al.* proposed the concept of category of mobility to reveal the benefits of natural processes by adapting natural transformation.[118]

Finally, it should be remarked that the aim of photonic approaches differs depending on the underlying physical mechanisms. The ultrahigh bandwidth attributes of the light wave provides superior properties to the electrical implementations as observed in optical telecommunications. In this sense, the laser chaos-based decision making (Sec. III E 1) exploits such an ultrafast aspect of the light wave, which would be unachievable by means of an electrical counterpart. The single-photon decision maker (Sec. III E 1) utilizes the wave-particle duality of light quanta; here, we see an exploitation of the quantum nature of light for decision making. Furthermore, entangled photons have been recently demonstrated for decision making[119] wherein quantum superpositions of photon states are utilized for making decisions. The significance of light-based approaches over the other platforms will be then determined by factors such as energy efficiencies, practical costs of devices, etc.. It has been shown that the energy efficiency of optical near-field-mediated energy transfer is 10^4 times more efficient than electron transport.[101] Also, the merit of an optical approach will be found in combinations with other optical processing systems, such as the fusion of optical decision making and photonic reservoir computing demonstrated in Ref. 97.

F. Compressed sampling accelerator

The traditional approach to digital acquisition of information samples an analog signal at or above the Nyquist rate, followed by the compression before storing bits. For a continuous-amplitude sequence of the signal either in time domain or space, the Nyquist rate is necessary to attain zero-distortion reconstruction of the input information in the case with constraint-free quantization. On the other hand, the minimal sampling rate for attaining the minimal distortion achievable in the presence of quantization constraint is usually below the Nyquist rate.[120] Compressed sampling (CS) realizes the minimal sampling rate for attaining the minimal distortion achievable within the quantization constraint is usually below the Nyquist rate.[121,122] CS relies on the fact that many types of information have a property called sparseness in the transformation process. The problem of CS is treated as an underdetermined linear system with prior information that the true solution is sparse, and the sparse signal vector can be reconstructed based on l_1-norm optimization.

1. Principle of operation

CS combines sampling and compression to obtain the measured vector $\boldsymbol{y}$ into a single linear operation by M-time performing inner products between a test matrix with random elements $\boldsymbol{\Phi}$ and the vector representation $\boldsymbol{x}$ of N-element data of the input information, given by

$$\boldsymbol{y} = \boldsymbol{\Phi x}, \tag{10}$$

where $\boldsymbol{\Phi}$ is $M \times N$ matrix and $M < N$. K represents the sparseness, counting the nonzero elements in N-element as shown in Fig. 23. The compression ratio is given by M/N. As Eq. (10) for the case with

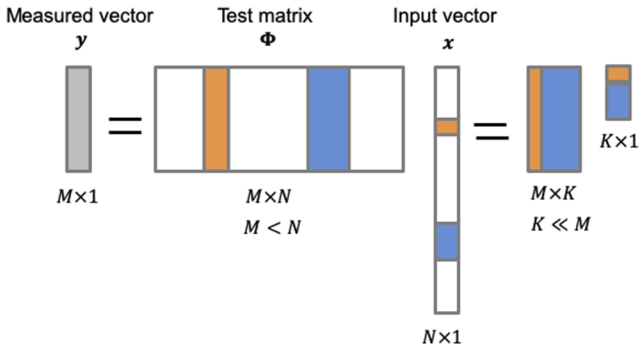

FIG. 23. Inner products between a test matrix $\boldsymbol{\Phi}$ and the vector representation $\boldsymbol{x}$.

$M < N$ is an ill-posed problem, there exists infinite number of solutions. However, if the random matrix $\boldsymbol{\Phi}$ satisfies with the restricted isometry property (RIP) that any selection of $2K$ columns of the random matrix $\boldsymbol{\Phi}$ is close to an orthobasis, the K-sparse vector $\hat{x}$ can be reconstructed by solving a convex optimization problem of l_1 norm given by[123]

$$\hat{\boldsymbol{x}} = \arg\min_{y=\Phi x} \|\boldsymbol{x}\|_1 \text{ subject to } \boldsymbol{\Phi x} = \boldsymbol{y}, \tag{11}$$

and the number of measurements M is in the order of

$$M = O(K \log(N/K)). \tag{12}$$

Notable advantage of CS implementation is that it enables capturing either image or time-serial data with a single photodetector, the so-called single-pixel camera, without using a spatially resolving detector such as a CMOS imager. This facilitates the operation across a broader spectral range where Si is blind, and even in the terahertz regime, CS-based imaging was demonstrated.[123] The single-pixel camera utilizes a digital micromirror device (DMD) as the spatial light modulator to realize the test matrix $\boldsymbol{\Phi}$ with random elements, and it demonstrated for the first time well-resolved imaging for $N = 256 \times 256$ photo with the compression ratio $M/N = 10\%$ –20%.[124] The frame rate of DMD device is around 4 kHz, which is much faster than 30 Hz of the CMOS imager, but the resolution or number of pixels is 1 Mega, much smaller than the CMOS imager.

2. Imagefree ghost imaging

Recently, image-free morphology-based cytometry, the so-called ghost imaging technique in Fig. 24, has been developed for cytometry, in which cell classification is enabled by machine-learning directly on the measured vector $\boldsymbol{y}$ without image reconstruction of vector $\hat{\boldsymbol{x}}$.[125] For applications such as classifications without need of rigid reconstruction of input information, this ghost imaging technique could be promising to speed up the processing.

A comparison with conventional NN will validate the precision of the ghost imaging technique. For instance, human activity recognition by both methods is numerically compared. The human activities including six motions such as walking, walking upstairs, walking downstairs, sitting, standing, and lying were detected by the accelerator of a smartphone. The captured temporal waveforms along each x-, y-, and z-axes for 2.56 s were sampled at the rate of

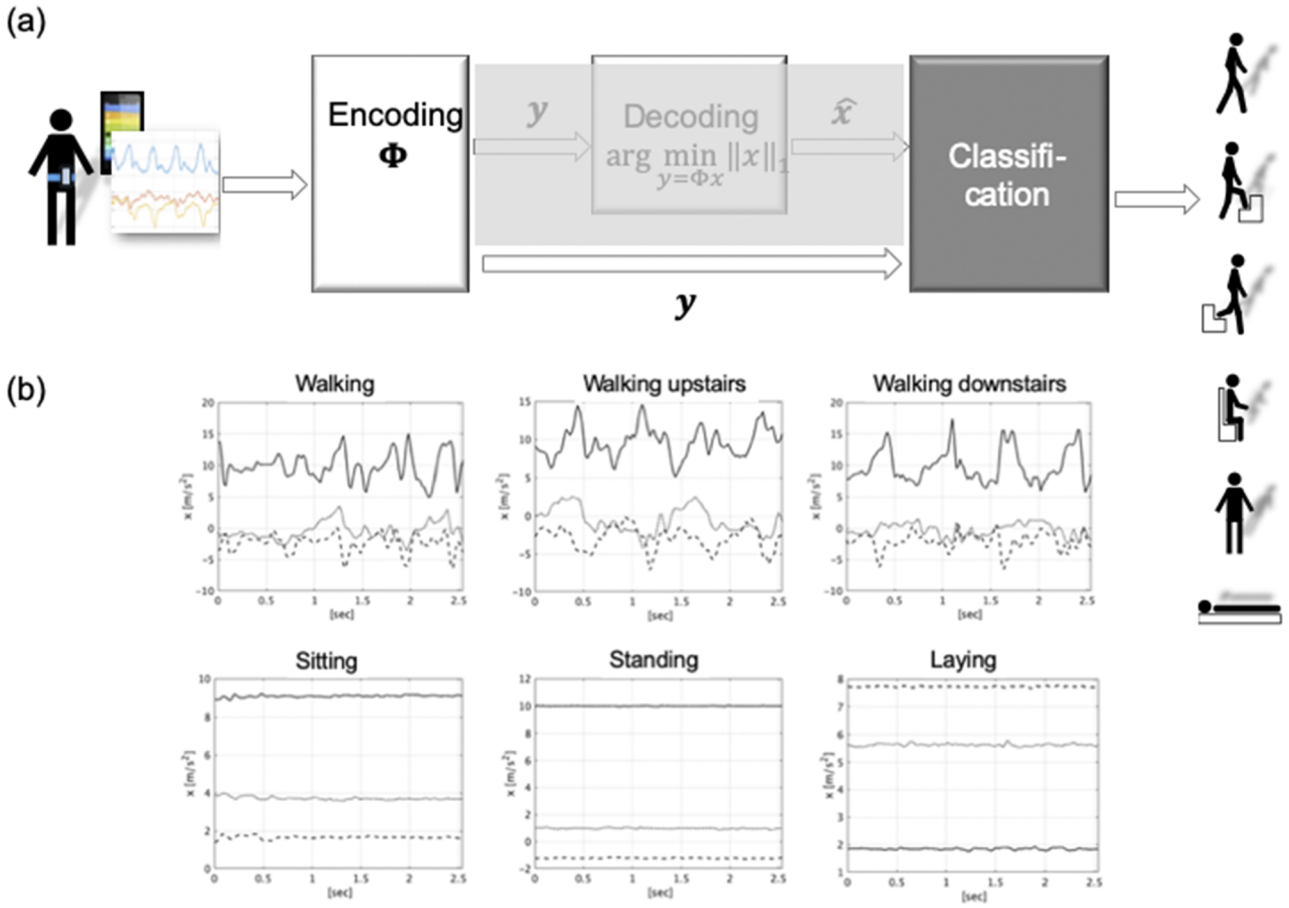

FIG. 24. (a) Conventional compressed sensing vs ghost imaging without image reconstruction and (b) temporal waveforms of six human motions of walking, walking upstairs, walking downstairs, sitting, standing, and lying. Visualized the data from Ref. 126.

20 ms into 128-sample. The data sets of the training and the testing include 7352 and 2947, respectively.[126] The three-layer NN consisting of the 384 (=128 × 3)-neuron input layer, 20-neuron hidden layer, and 6-neuron output layer is used for the simulation. The compression ratio is 50% that means that 128-sample is compressed into 64-sample. The ghost imaging performs fairly good, showing the precision of 81.3%, although it is slightly outperformed by the NN result of the precision of 94.7%

3. Challenges and future perspective

CMOS and CCD imager works only in the visible spectrum in 300–1100 nm. However, a high-density detector array is unfeasible at longer wavelength such as terahertz regime (300 GHz–10 THz or λ = 1 mm–30 μm) due to a fundamental lack of suitable materials for the imaging devices. CS enables a single-pixel sensing in a wide spectral range, including metamaterial absorbers in the THz regime,[123,127] a photomultiplier in the UV to infrared region, and a microwave via EO conversion using a light modulator.[128] Therefore, it paves a way for versatile light detection in a simple configuration with a desired sensitivity and spectral region.

CS can be extensively applied to communications, medicine, biology, astronomy, etc. In communications, wireless channel estimation, wireless sensor network, network tomography, cognitive radio, array signal processing, multiple access schemes, and networked control.[129] It is also applied to a wide variety of imaging such as photoacoustic imaging, optical coherence tomography (OCT), electron microscopy, and fluorescence microscopy. Very recently, imaging of the shadow of a black hole observed by the Event Horizon Telescope's (EHT's) global network of synchronized radio observatories was reported, in which CS played a key role for the reconstruction of raw data.[130] This event will further prompt applications of CS to the so far unexplored sectors in the near future.

Finally, an emphasis is put on a combination of a smart mobile tablet attached with a photonic sensor module and AI in the edge/fog computing over 5G URLLC link in Fig. 6. This enables real-time operation within the response of less than 1 ms. CS-based sensing facilitates data acquisition of a much smaller volume than conventional sampling and compressing of data. Hence, preprocessing of the data can be performed with limited computation resource of a smart tablet, followed by transporting only the core data to the fog/edge for the data analytics based upon AI.

IV. CONCLUSION

There has been a growing demand for computing power in the IoT-CPS embedded society. The way people use computers has been radically changing to cloud computing without the ownership of computing resources. Cloud computing provides users with various types of services such as XaaS. Particularly, edge/fog computing in microdatacenters, deployed as close as possible where the events happen, enables time-sensitive applications such as autonomous vehicle and augmented reality/virtual reality AR/VR and industrial robots. It realizes mobile computing everywhere over low-latency 5G wireless link between a smart tablet of the user and edge/fog computing resource.

On the other hand, digital computing is facing a bottleneck as Moore's law has been coming to the end, and parallel computation exhibits its own limitation, although it continues sustaining the throughput of computation at the expense of energy consumption. One of the solution paths to overcome the bottleneck from the hardware approach is post-Moore optoelectronic circuit technologies. Another approach is hardware accelerators such as GPU and FPGA placed at the front end of a digital computer.

We have introduced a photonic accelerator (PAXEL), which is a special class of processor placed at the front end of digital computer. It is distinct from electronic accelerators in that the target information is optically sensed and processed. It is optimized to perform a specific function but does so faster with less power consumption than the electronic general-purpose processor. We have reviewed an array of PAXEL architectures and applications, including ANNs, reservoir computing, pass-gate logic, decision making, and compressed sensing. Promising computing architectures for PAXEL have been presented, including neuromorphic computing, reservoir computing, pass-gate logic, reinforcement learning, and compressed sensing. We have assessed the potential advantages and challenges for each of these PAXEL approaches to highlight the scope for future work toward practical implementation. We hope that this article prompts pioneering new frontiers of photonics for data processing and the PAXEL eventually becomes an intelligent mobile tool in daily life.

ACKNOWLEDGMENTS

This research was supported by the CREST program "Advanced core technology for creation and practical utilization of innovative properties and functions based upon optics and photonics," funded by the Japanese Science and Technology Agency. The authors would like to thank the reviewers for their detailed comments and insightful suggestions for the revisions.

REFERENCES

[1] E. Mäkiö-Marusik, B. Ahmad, R. Harrison, J. Mäkiö, and A. Walter, "Competences of cyber physical systems engineers—Survey results," in 2018 IEEE Industrial Cyber-Physical Systems, St. Petersburg, May 2018.

[2] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in MCC2012 Workshop on Mobile Cloud Computing, Helsinki, Finland, August 2012.

[3] L. Peterson, A. Al-Shabibi, T. Anshutz, S. Baker, A. Bavier, S. Das, J. Hart, G. Palukar, and W. Snow, "Central office Re-architected as a data center," IEEE Commun. Mag. **54**(10), 96–101 (2016).

[4] See https://1.ieee802.org/tsn/#Published_TSN_Standards for IEEE 802.1 Time-Sensitive Networking (TSN) Task Group.

[5] R. H. Dennard, "Evolution of the MOSFET dynamic RAM—A personal view," IEEE Trans. Electron Devices **31**(11), 1549–1555 (1984).

[6] R. H. Dennard, F. H. Gaesslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, "Design of ion-implanted MOSFET's with very small physical dimensions," IEEE J. Solid-State Circuits **9**(5), 256–268 (1974).

[7] G. E. Moore, "Cramming more components onto integrated circuits," Proc. IEEE **86**(1), 82 (1998).

[8] *History of Processor Performance* (Columbia University, 2012), http://www.cs.columbia.edu/~sedwards/classes/2012/3827-spring/advanced-arch-2011.pdf.

[9] M. D. Hill and M. R. Marty, "Amdahl's law in the multicore era," Computer **41**(7), 33–38 (2008).

[10] J. Koomey, S. Berard, M. Sanchez, and H. Wong, "Implications of historical trends in the electrical efficiency of computing," IEEE Ann. Hist. Comput. **33**(3), 46–54 (2011).

[11] B. Marr, B. Degnan, P. Hasler, and D. Anderson, "Scaling energy per operation via an asynchronous pipeline," IEEE Trans. Very Large Scale Integr. Syst. **21**(1), 147–151 (2013).

[12] A. Sano, T. Kobayashi, S. Yamanaka, A. Matsuura, H. Kawakami, Y. Miyamoto, K. Ishihara, and H. Masuda, "102.3-Tb/s (224 x 548-Gb/s) C- and extended L-band all-Raman transmission over 240 km using PDM-64QAM single carrier FDM with digital pilot tone," in *OFC2012* (OSA, 2012), p. PDP5C.3.

[13] D. Soma, Y. Wakayama, S. Beppu, S. Sumita, T. Tsuritani, T. Hayashi, T. Nagashima, M. Suzuki, M. Yoshida, K. Kasai, M. Nakazawa, H. Takahashi, K. Igarashi, I. Morita, and M. Suzuki, "10.16-peta-B/s dense SDM/WDM transmission over 6-mode 19-core fiber across the C+L band," J. Lightwave Technol. **36**(6), 1362–1368 (2018).

[14] R. K. Cavin III, P. Lugli, and V. V. Zhirnov, "Science and engineering beyond Moore's law," Proc. IEEE **100**, 1720–1749 (2012).

[15] K. Yamada, "Silicon photonics for a post-Moore era," PIC Magazine, December 2016, https://picmagazine.net/article/101212/Silicon_Photonics_For_A_Post-Moore_Era/feature.

[16] International Technology Roadmap for Semiconductors (ITRSs) 2.0, 2015 Ed.

[17] See http://www.aimphotonics.com for information about AIM Photonics.

[18] B. Jalali and A. Mahjoubfar, "Tailoring wideband signals with a photonic hardware accelerator," Proc. IEEE **103**(7), 1071–1086 (2015).

[19] Recommendation ITU-R M.2083-0, IMT Vision—Frame-work and overall objectives of the future development of IMT for 2020 and beyond, September 2015.

[20] E. Mcleod and A. Ozcan, "Microscopy without lenses," Phys. Today **70**(9), 50–56 (2017).

[21] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, and O. Temam, "DaDianNao: A machine-learning supercomputer," in *International Symposium on Microarchitecture* (IEEE/ACM, 2014), pp. 609–622.

[22] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P. L. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon, "In-datacenter performance analysis of a tensor processing unit," in *International Symposium on Computer Architecture* (IEEE/ACM, 2017), pp. 1–17.

[23] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "ISAAC: A convolutional neural network accelerator with *in situ* analog arithmetic in crossbars," in *International Symposium on Computer Architecture* (ACM, 2016), pp. 14–26.

[24] 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," Nat. Photonics **11**(7), 441 (2017).

[25]S. Cartwright, "New optical matrix–vector multiplier," Appl. Opt. **23**(11), 1683–1684 (1984).
[26]M. Gruber, J. Jahns, and S. Sinzinger, "Planar-integrated optical vector-matrix multiplier," Appl. Opt. **39**(29), 5367–5373 (2000).
[27]A. N. Tait, T. Ferreira de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastr, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**(1), 7430 (2017).
[28]N. Janosik, Q. Cheng, M. Glick, Y. Huang, and K. Bergman, "High-resolution silicon microring based architecture for optical matrix multiplication," in *Conference on Lasers and Electro-Optics, OSA Technical Digest* (OSA, 2019), p. SM2J.3.
[29]M. Bahadori, M. Nikdast, S. Rumley, L. Yuan Dai, N. Janosik, T. V. Vaerenbergh, A. Gazman, Q. Cheng, R. Polster, and K. Bergman, "Design space exploration of microring resonators in silicon photonic interconnects: Impact of the ring curvature," J. Lightwave Technol. **36**(13), 2767 (2018).
[30]M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," Phys. Rev. Lett. **73**(1), 58–61 (1994).
[31]W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, "Optimal design for universal multiport interferometers," Optica **3**(12), 1460–1465 (2016).
[32]S. Kawakami, T. Ono, M. Notomi, and K. Inoue, "Evaluation platform for a nanophotonic neural network accelerator (in Japanese)," IEICE Trans. **J102-A**(6), 182–193 (2019).
[33]S. Beamer, C. Sun, Y. Kwon, A. Joshi, C. Batten, V. Stojanović, and K. Asanović, "Re-architecting DRAM memory systems with monolithically integrated silicon photonics," in *International Symposium on Computer Architecture* (ISCA, 2010), pp. 129–140.
[34]W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," Neural Comput. **14**, 2531–2560 (2002).
[35]H. Jaeger and H. Hass, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," Science **304**, 78–80 (2004).
[36]L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, "Information processing using a single dynamical node as complex system," Nat. Commun. **2**, 468 (2011).
[37]L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer, "Photonic information processing beyond turing: An optoelectronic implementation of reservoir computing," Opt. Express **20**(3), 3241–3249 (2012).
[38]Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, "Optoelectronic reservoir computing," Sci. Rep. **2**, 287 (2012).
[39]F. Duport, B. Schneider, A. Smerieri, A. M. Haelterman, and S. Masser, "All-optical reservoir computing," Opt. Express **20**(20), 22783–22795 (2012).
[40]D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," Nat. Commun. **4**, 1364 (2013).
[41]J. Bueno, D. Brunner, M. C. Soriano, and I. Fischer, "Conditions for reservoir computing performance using semiconductor lasers with delayed optical feedback," Opt. Express **25**(3), 2401–2412 (2017).
[42]J. Nakayama, K. Kanno, and A. Uchida, "Laser dynamical reservoir computing with consistency: An approach of a chaos mask signal," Opt. Express **24**(8), 8679–8692 (2016).
[43]Y. Kuriki, J. Nakayama, K. Takano, and A. Uchida, "Impact of input mask signals on delay-based photonic reservoir computing with semiconductor lasers," Opt. Express **26**(5), 5777–5788 (2018).
[44]J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner, "Reinforcement learning in a large-scale photonic recurrent neural network," Optica **5**(6), 756–760 (2018).
[45]A. S. Weigend and N. A. Gershenfeld, "Results of the time series prediction competition at the Santa Fe Institute," in *IEEE International Conference on Neural Networks* (IEEE, 1993), Vol. 3, pp. 1786–1793.
[46]A. Uchida, R. McAllister, and R. Roy, "Consistency of nonlinear system response to complex drive signals," Phys. Rev. Lett. **93**, 244102 (2004).
[47]R. Lang and K. Kobayashi, "External optical feedback effects on semiconductor injection laser properties," IEEE J. Quantum Electron. **16**(3), 347–355 (1980).
[48]A. Uchida, *Optical Communication with Chaotic Lasers, Applications of Nonlinear Dynamics and Synchronization* (Wiley-VCH, 2012).
[49]F. Laporte, A. Katumba, J. Dambre, and P. Bienstman, "Numerical demonstration of neuromorphic computing with photonic crystal cavities," Opt. Express **26**(7), 7955–7964 (2018).
[50]D. Brunner and I. Fischer, "Reconfigurable semiconductor laser networks based on diffractive coupling," Opt. Lett. **40**(16), 3854–3857 (2015).
[51]S. Kreinberg, X. Porte, D. Schicke, B. Lingnau, C. Schneider, S. Höfling, I. Kanter, K. Lüdge, and S. Reitzenstein, "Mutual coupling and synchronization of optically coupled quantum-dot micropillar lasers at ultra-low light levels," Nat. Commun. **10**, 1539 (2019).
[52]P. Antonik, M. Haelterman, and S. Masser, "Brain-inspired photonic signal processor for generating periodic patterns and emulating chaotic systems," Phys. Rev. Appl. **7**, 054014 (2017).
[53]A. Akrout, A. Bouwens, F. Duport, Q. Vinckier, M. Haelterman, and S. Masser, "Parallel photonic reservoir computing using frequency multiplexing of neurons," e-print arXiv:1612.08606v1 (2016).
[54]K. Takano, C. Sugano, M. Inubushi, K. Yoshimura, S. Sunada, K. Kanno, and A. Uchida, "Compact reservoir computing with a photonic integrated circuit," Opt. Express **26**(22), 29424–29439 (2018).
[55]K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, "Experimental demonstration of reservoir computing on a silicon photonics chip," Nat. Commun. **5**, 3541 (2014).
[56]A. Argyris, J. Bueno, and I. Fischer, "Photonic machine learning implementation for signal recovery in optical communications," Sci. Rep. **8**, 8487 (2018).
[57]A. Argyris, J. Bueno, and I. Fischer, "PAM-4 transmission at 1550 nm using photonic reservoir computing post-processing," IEEE Access **7**, 37017–37025 (2019).
[58]C. Gallicchio, A. Micheli, and L. Pedrelli, "Deep reservoir computing: A critical experimental analysis," Neurocomputing **268**, 87–99 (2017).
[59]D. Dhang, S. A. Hasnain, and R. Mahapatra, "MReC: A multilayer photonic reservoir computing architecture," in 20th International Symposium on Quality Electronic Design (ISQED), 2019.
[60]M. Nakajima, S. Konisho, K. Tanaka, and T. Hashimoto, "Deep reservoir computing using delay-based optical nonlinear oscillator," in Cognitive Computing Conference, 2018.
[61]T. Okumura, M. Tai, and M. Ando, "Optical implementation of reservoir computing for fast integrative analysis in sensor array processing," in *Proceedings of 2017 International Symposium on Nonlinear Theory and Its Applications (NOLTA 2017)* (IEICE, 2017), pp. 256–259.
[62]S. Sunada, K. Arai, and A. Uchida, "Wave dynamical reservoir computing at a microscale," in *Proceedings of 2018 International Symposium on Nonlinear Theory and Its Applications (NOLTA 2018)* (IEICE, 2018), pp. 154–155.
[63]International Roadmap for Devices and Systems (IRDS) 2017 Ed., 2017, https://irds.ieee.org/roadmap-2017.
[64]S. Werner, J. Navaridas, and M. Luján, "A survey on optical network-on-chip architectures," ACM Comput. Surv. **50**(6), 1–37 (2017).
[65]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," in *Proceedings of Interconnect Technology Conference/Advanced Metallization Conference* (IEEE, 2014), pp. 345–348.
[66]D. A. B. Miller, "Attojoule optoelectronics for low-energy information processing and communications," J. Lightwave Technol. **35**(3), 346–396 (2017).
[67]C. Debaes *et al.*, "Receiver-less optical clock injection for clock distribution networks," IEEE J. Sel. Top. Quantum Electron. **9**(2), 400–409 (2003).
[68]M. Notomi, K. Nozaki, A. Shinya, S. Matsuo, and E. Kuramochi, "Toward fJ/bit optical communication in a chip," Opt. Commun. **314**, 3–17 (2014).
[69]C. Haffner, W. Heni, Y. Fedoryshyn, J. Niegemann, A. Melikyan, D. L. Elder, B. Baeuerle, Y. Salamin, A. Josten, U. Koch, C. Hoessbacher, F. Ducry, L. Juchli, A. Emboras, D. Hillerkuss, M. Kohl, L. R. Dalton, C. Hafner, and J. Leuthold, "All-plasmonic Mach–Zehnder modulator enabling optical high-speed communication at the microscale," Nat. Photonics **9**, 525–528 (2015).

[70]A. Fujiwara, N. M. Zimmerman, Y. Ono, and Y. Takahashi, "Current quantization due to single-electron transfer in Si-wire charge-coupled devices," Appl. Phys. Lett. **84**, 1323–1325 (2004).
[71]S. S. Agashe, K. T. Shiu, and S. R. Forrest, "Integratable high linearity compact waveguide coupled tapered InGaAsP photodetectors," IEEE J. Quantum Electron. **43**, 597–606 (2007).
[72]H. Chen, P. Verheyen, P. De Heyn, G. Lepage, J. De Coster, S. Balakrishnan, P. Absil, W. Yao, L. Shen, G. Roelkens, and J. Van Campenhout, "-1 V bias 67 GHz bandwidth Si-contacted germanium waveguide p-i-n photodetector for optical links at 56 Gbps and beyond," Opt. Express **24**, 4622–4631 (2016).
[73]S. Lischke, D. Knoll, C. Mai, L. Zimmermann, A. Peczek, M. Kroh, A. Trusch, E. Krune, K. Voigt, and A. Mai, "High bandwidth, high responsivity waveguide-coupled germanium p-i-n photodiode," Opt. Express **23**, 27213–27220 (2015).
[74]V. J. Sorger, N. D. Lanzillotti-Kimura, R. M. Ma, and X. Zhang, "Ultra-compact silicon nanophotonic modulator with broadband response," Nanophotonics **1**, 17–22 (2012).
[75]K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, M. Ono, A. Shakoor, E. Kuramochi, and M. Notomi, "Photonic-crystal nano-photodetector with ultrasmall capacitance for on-chip light-to-voltage conversion without an amplifier," Optica **3**(5), 483–492 (2016).
[76]J. Joannopoulos, S. G. Johnson, R. Meade, and J. Winn, *Photonic Crystals, Molding the Flow of Light*, 2nd ed. (Princeton University Press, 2007).
[77]M. Notomi, "Manipulating light with strongly modulated photonic crystals," Rep. Prog. Phys. **73**, 096501 (2010).
[78]K. Nozaki, S. Matsuo, A. Shinya, and M. Notomi, "Amplifier-free bias-free receiver based on low-capacitance nanophotodetector," IEEE J. Sel. Top. Quantum Electron. **24**(2), 4900111 (2018).
[79]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**, 046101 (2018).
[80]K. Nozaki, A. Shakoor, S. Matsuo, T. Fujii, K. Takeda, A. Shinya, E. Kuramochi, and M. Notomi, "Ultralow-energy electro-absorption modulator consisting of InGaAsP-embedded photonic-crystal waveguide," APL Photonics **2**, 056105 (2017).
[81]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," Nat. Photonics **13**, 454 (2019).
[82]J. Hardy and J. Shamir, "Optics inspired logic architecture," Opt. Express **15**(1), 150–165 (2007).
[83]S. Lin, Y. Ishikawa, and K. Wada, "Demonstration of optical computing logics based on binary decision diagram," Opt. Express **20**(2), 1378–1384 (2012).
[84]Q. Xu and R. Sorel, "Reconfigurable optical directed-logic circuits using microresonator-based optical switches," Opt. Express **19**(6), 5244–5259 (2011).
[85]T. Ishihara, A. Shinya, K. Inoue, K. Nozaki, and M. Notomi, "An integrated nanophotonic parallel adder," ACM J. Emerg. Technol. Comput. Syst. **14**(2), 26 (2018).
[86]S. Kita, K. Nozaki, K. Takata, A. Shinya, and M. Notomi, "Silicon linear optical logic gates for low-latency computing," in CLEO 2018, San Jose, USA, 2018, paper SF1A.2.
[87]L. Lai, H. El Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: Exploration, exploitation, and competition," IEEE Trans. Mobile Comput. **10**(2), 239–253 (2011).
[88]K. Kuroda, H. Kato, S.-J. Kim, M. Naruse, and M. Hasegawa, "Improving throughput using multi-armed bandit algorithm for wireless LANs," Nonlinear Theory Appl., IEICE **9**, 74–81 (2018).
[89]D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," Nature **550**, 354 (2017).
[90]R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (The MIT Press, Massachusetts, 1998).
[91]N. Daw, J. O'Doherty, P. Dayan, B. Seymour, and R. Dolan, "Cortical substrates for exploratory decisions in humans," Nature **441**, 876–879 (2006).
[92]H. Robbins, "Some aspects of the sequential design of experiments," Bull. Am. Math. Soc. **58**, 527–535 (1952).
[93]P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multi-armed bandit problem," Mach. Learn. **47**, 235–256 (2002).
[94]J. Backus, "Can programming be liberated from the von Neumann style?: A functional style and its algebra of programs," Commun. ACM **21**(8), 613–641 (1978).
[95]T. Inagaki, Y. Haribara, K. Igarashi, T. Sonobe, S. Tamate, T. Honjo, A. Marandi, P. L. McMahon, T. Umeki, K. Enbutsu, O. Tadanaga, H. Takenouchi, K. Aihara, K. Kawarabayashi, K. Inoue, S. Utsunomiya, and H. Takesue, "A coherent ising machine for 2000-node optimization problems," Science **354**(6312), 603–606 (2016).
[96]J. H. Han, F. Boeuf, J. Fujikata, S. Takahashi, S. Takagi, and M. Takenaka, "Efficient low-loss InGaAsP/Si hybrid MOS optical modulator," Nat. Photonics **11**(8), 486–490 (2017).
[97]K. Kanno, M. Naruese, and A. Uchida, "Optical reservoir computing with combination of reinforcement learning," in *The 79th Japan Society of Applied Physics Autumn Meeting* (JSAP, 2018), pp. 03–139.
[98]M. Naruse, M. Berthel, A. Drezet, S. Huant, M. Aono, H. Hori, and S.-J. Kim, "Single-photon decision maker," Sci. Rep. **5**, 13253 (2015).
[99]M. Naruse, M. Berthel, A. Drezet, S. Huant, H. Hori, and S.-J. Kim, "Single photon in hierarchical architecture for physical decision making: Photon intelligence," ACS Photonics **3**, 2505–2514 (2016).
[100]M. Berthel, O. Mollet, G. Dantelle, T. Gacoin, S. Huant, and A. Drezet, "Photophysics of single nitrogen-vacancy centers in diamond nanocrystals," Phys. Rev. B **91**, 035308 (2015).
[101]M. Naruse, N. Tate, M. Aono, and M. Ohtsu, "Information physics fundamentals of nanophotonics," Rep. Prog. Phys. **76**, 056401 (2013).
[102]S.-J. Kim, M. Naruse, M. Aono, M. Ohtsu, and M. Hara, "Decision maker based on nanoscale photo-excitation transfer," Sci. Rep. **3**, 2370 (2013).
[103]M. Naruse, W. Nomura, M. Aono, M. Ohtsu, Y. Sonnefraud, A. Drezet, S. Huant, and S.-J. Kim, "Decision making based on optical excitation transfer via near-field interactions between quantum dots," J. Appl. Phys. **116**, 154303 (2014).
[104]J. Ohtsubo, *Semiconductor Lasers: Stability, Instability and Chaos* (Springer, Berlin, 2012).
[105]A. Uchida, *Optical Communication with Chaotic Lasers: Applications of Nonlinear Dynamics and Synchronization* (Wiley-VCH, Weinheim, 2012).
[106]A. Uchida, K. Amano, M. Inoue, K. Hirano, S. Naito, H. Someya, I. Oowada, T. Kurashige, M. Shiki, S. Yoshimori, K. Yoshimura, and P. Davis, "Fast physical random bit generation with chaotic semiconductor lasers," Nat. Photonics **2**, 728–732 (2008).
[107]A. Argyris, S. Deligiannidis, E. Pikasis, A. Bogris, and D. Syvridis, "Implementation of 140 Gb/s true random bit generator based on a chaotic photonic integrated circuit," Opt. Exp. **18**, 18763–18768 (2010).
[108]M. Naruse, Y. Terashima, A. Uchida, and S.-J. Kim, "Ultrafast photonic reinforcement learning based on laser chaos," Sci. Rep. **7**, 8772 (2017).
[109]R. F. Fox, I. R. Gatland, R. Roy, and G. Vemuri, "Fast, accurate algorithm for numerical simulation of exponentially correlated colored noise," Phys. Rev. A **38**, 5938–5940 (1988).
[110]M. Naruse, T. Mihana, H. Hori, H. Saigo, K. Okamura, M. Hasegawa, and A. Uchida, "Scalable photonic reinforcement learning by time-division multiplexing of laser chaos," Sci. Rep. **8**, 10890 (2018).
[111]T. Mihana, Y. Mitsui, M. Naruse, and A. Uchida, "Decision making using lag synchronization of chaos in mutually-coupled semiconductor lasers," in *2018 International Symposium on Nonlinear Theory and Its Applications* (IEICE, 2018), pp. 215–218.
[112]R. Homma, S. Kochi, T. Niiyama, A. Uchida, M. Naruse, and S. Sunada, "Optical decision making with a semiconductor ring laser," in *2018 International Symposium on Nonlinear Theory and Its Applications* (IEICE, 2018), pp. 211–214.
[113]R. Nakagomi, K. Uchiyama, H. Suzui, E. Hatano, K. Uchida, M. Naruse, and H. Hori, "Nanometre-scale pattern formation on the surface of a photochromic crystal by optical near-field induced photoisomerization," Sci. Rep. **8**, 14468 (2018).
[114]T. Mihana, Y. Terashima, M. Naruse, S.-J. Kim, and A. Uchida, "Memory effect on adaptive decision making with a chaotic semiconductor laser," Complexity **2018**, 4318127.

[115]K. Katayama, K. Takano, S. Amakawa, S. Hara, A. Kasamatsu, K. Mizuno, and K. Takahashi, "300 GHz CMOS transmitter with 32-QAM 17.5 Gb/s/ch capability over six channels," IEEE J. Solid-State Circuits **51**(12), 3037–3048 (2016).
[116]M. Naruse, S.-J. Kim, M. Aono, M. Berthel, A. Drezet, S. Huant, and H. Hori, "Category theoretic analysis of photon-based decision making," Int. J. Inf. Technol. Decis. Making **17**(5), 1305–1333 (2018).
[117]M. Naruse, E. Yamamoto, T. Nakao, T. Akimoto, H. Saigo, K. Okamura, I. Ojima, G. Northoff, and H. Hori, "Why is the environment important for decision making? Local reservoir model for choice-based learning," PLoS One **13**(10), e0205161 (2018).
[118]H. Saigo, M. Naruse, K. Okamura, H. Hori, and I. Ojima, "Analysis of soft robotics based on the concept of category of mobility," Complexity **2019**, 1490541.
[119]M. Naruse, N. Chauvet, D. Jegouso, B. Boulanger, H. Saigo, K. Okamura, H. Hori, A. Drezet, S. Huant, and G. Bachelier, "Entangled photons for competitive multi-armed bandit problem: Achievement of maximum social reward, equality, and deception prevention," e-print arXiv:1804.04316.
[120]A. Kipnis, Y. Eldar, and A. Goldsmith, "Analog-to-digital compression: A new paradigm for converting signals to bits," IEEE Signal Process. Mag. **35**, 16–39 (2018).
[121]D. Donoho, "Compressed sensing," IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006).
[122]R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," Constructive Approximation **28**, 253–263 (2008).
[123]W. L. Chan, M. L. Moravec, R. G. Baraniuk, and D. M. Mittleman, "Terahertz imaging with compressed sensing and phase retrieval," Opt. Lett. **33**(9), 974–976 (2008).
[124]M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling—Building simpler, smaller, and less-expensive digital cameras," in *IEEE Signal Processing Magazine* (IEEE, 2008), pp. 83–919.
[125]S. Ohta, R. Horisaki, Y. Kawamura, M. Ugawa, I. Sato, K. Hashimoto, R. Kamesawa, K. Setoyama, S. Yamaguchi, K. Fujiu, K. Waki, and H. Noji, "Ghost cytometry," Science **360**(6394), 1246–1251 (2018).
[126]D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International Workshop of Ambient Assisted Living (IWAAL2012)* (Vitoria-Gasteiz, Spain, December 2012).
[127]C. M. Watts, D. Shrekenhamer, J. Montoya, G. Lipworth, J. Hunt, T. Sleasman, S. Krishna, D. R. Smith, and W. J. Padilla, "Terahertz compressive imaging with metamaterial spatial light modulators," Nat. Photonics **8**, 605–609 (2014).
[128]H. Chi, Y. Chen, Y. Mei, X. Jin, S. Zheng, and X. Zhang, "Microwave spectrum sensing based on photonic time stretch and compressive sampling," Opt. Lett. **38**(2), 136–138 (2013).
[129]K. Hayashi, M. Nagahara, and T. Tanaka, "A user's guide to compressed sensing for communications systems," IEICE Trans. Commun. **E96-B**(3), 685–712 (2013).
[130]D. Castelvecchi, "Black hole pictured for first time-in spectacular detail," Nature **568**, 284–285 (2019).