Research article

Pascal Stark, Folkert Horst, Roger Dangel, Jonas Weiss and Bert Jan Offrein*

Opportunities for integrated photonic neural networks

https://doi.org/10.1515/nanoph-2020-0297
Received May 19, 2020; accepted July 20, 2020; published online August 10, 2020

Abstract: Photonics offers exciting opportunities for neuromorphic computing. This paper specifically reviews the prospects of integrated optical solutions for accelerating inference and training of artificial neural networks. Calculating the synaptic function, thereof, is computationally very expensive and does not scale well on state-of-the-art computing platforms. Analog signal processing, using linear and nonlinear properties of integrated optical devices, offers a path toward substantially improving performance and power efficiency of these artificial intelligence workloads. The ability of integrated photonics to operate at very high speeds opens opportunities for time-critical real-time applications, while chip-level integration paves the way to cost-effective manufacturing and assembly.

Keywords: integrated optics; optical signal processing; photonic neural networks; photonic reservoir computing.

1 Introduction

Over the last two decades, the computing landscape has massively changed. The saturation of silicon technology scaling started to cripple Moore's law, and as a consequence, new architectures and integration schemes had to be developed to maintain the computing performance roadmaps. The emergence of ultrahigh bandwidth internet facilitated a new 'computing-as-a-service' model based on large flexible disaggregated systems in the cloud and enabled new applications and services like video streaming, social networks and data-driven business intelligence. The availability of large amounts of data from, and for, such services naturally created a desire to extract value from them. However, because a large part of those data is noisy, unstructured or incomplete, traditional, statistical methods have difficulties working properly. This refueled interest in trainable or even self-learning algorithms that were already of great scientific interest in the 1960s and 1990s [1]. The revival of neuromorphic computing had been triggered. Exploiting the now finally available computing power of silicon technology, large and complex brain-inspired architectures can be designed, optimized, and executed, and artificial intelligence (AI) has become a major area of R&D and an essential part of our daily life.

A big challenge on the path to ultimate brain-inspired systems is that the brain itself is not yet understood well enough to take it as a starting point [2]. This holds true at all levels, from the smallest building block to the overall architecture, interaction and memory models. Therefore, in today's AI systems, the so-far identified building blocks and architectures are loosely mapped onto a suitable technology platform, up to the extent that the term 'brain-inspired' may even be a large stretch. Silicon Complementary metal-oxide-semiconductor (CMOS) is the most advanced, highest performance, miniaturized, reliable and established one. Hence, it is used as the basis for almost all AI hardware implementations today. To overcome the fundamental memory bottlenecks of the von Neumann architecture, it became necessary to advance silicon CMOS toward novel architectures [3] and enhance its functionality in a 'more-than-Moore' approach. One key aspect of the latter is to deeply embed the basic neural network building blocks like the massively parallel synaptic interconnect layers and nonlinear activation functions in the platform foundation.

The main task in neuromorphic computing is calculating and optimizing the synaptic interconnects in a neural network, wherein the signals into the neurons are weighted and summed through many multiply-accumulate (MAC) operations. If we consider all synaptic connections between two network layers, this operation can finally be

***Corresponding author: Bert Jan Offrein**, IBM Research – Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland, E-mail: ofb@zurich.ibm.com. https://orcid.org/0000-0001-6082-0068
Pascal Stark, Folkert Horst, Roger Dangel and Jonas Weiss, IBM Research – Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland, E-mail: crk@zurich.ibm.com (P. Stark), fho@zurich.ibm.com (F. Horst), rda@zurich.ibm.com (R. Dangel), jwe@zurich.ibm.com (J. Weiss). https://orcid.org/0000-0002-7389-5592 (P. Stark)

formulated as one large vector-matrix multiplication. The computational cost of the latter scales with N^2, with N being the number of neurons of the neural network layer. To accelerate the computation of large numbers of these vector-matrix multiplications during training and inference of deep neural networks (DNNs), dedicated hardware accelerators were introduced. Examples include graphics processing units (GPUs) [4] or tensor processing units [5]. Such accelerators enable parallel and pipelined processing of MAC operations, fetching data from memory and writing back the results. This moving back and forth of data/results between the different memory locations and the actual computing engines constitutes the classical von Neumann bottleneck and is attributed to the bulk of the overall energy consumption (Figure 1). For out-of-order instruction executions, it is probably the limiting factor in the overall performance of the classical computing.

With increasing distance to the processing unit, memory access becomes increasingly more power expensive and through larger access latency slows down computing (Figure 1). Furthermore, the energy for performing arithmetic operations strongly depends on the required accuracy. To mitigate the massive power consumption of today's systems, two directions become clear. First, data must be kept as local as possible. Second, operations must be performed at the lowest accuracy feasible. Using accelerators like GPUs with copackaged memory is in line with this concept. However, though data are kept more local and processing is massively parallelized, processing and memory units are still separated as in a von Neumann architecture. The urgency of overcoming these power-driving mechanisms of today's systems was recently assessed [7]. Strubell et al. [7] show that training a state-of-the-art natural language processing neural network, for which 213-M parameters had to be optimized on a cloud data center using modern GPUs, requires around 200 kWh. Even by using partially renewable energy for running the cloud data center, this still translates into the estimated emission of 100 kg of CO_2 to train one neural network (Figure 2).

Consequently, an enhanced technology platform must

(1) overcome excessive data motion;
(2) reduce signal processing overhead;
(3) provide synaptic connections resembling the neural network architecture.

The processor-to-memory data exchange issue can be largely addressed by pursuing in-memory computing concepts, while the signal processing overhead is reduced by applying analog signal processing. In the electrical domain, fascinating new concepts are emerging adhering to the concepts listed above. For example, the use of computational memory based on memristive devices enables to perform MAC operations, in-place (or in-memory). The MAC operation is performed in the analog domain using memristive devices by exploiting Ohm's law (multiply operation) and Kirchhoff's law (accumulate operation). The input signal is applied as a voltage across a conductor, and the resulting current is the multiplication of the voltage and the conductance. Combining the currents from multiple individual conductors leads to the accumulated current. Each memristive device represents a synaptic weight. Recently, impressive demonstrations of neural network operations employing electrical crossbar technology for calculating the synaptic interconnect were achieved [10]. The memristive devices can be integrated in the back-end-of-the-line of CMOS technology and are often implemented based on phase-change materials [11] or metal oxide resistive memory (OxRAM) devices [12]. The direct cointegration with CMOS and, hence, high-density implementation is an important aspect of this technology. Challenges remain in setting the resistance to the desired value and the retention thereof, as required for inference, while a well-controllable change of the resistance is important for efficient neural network training [13].

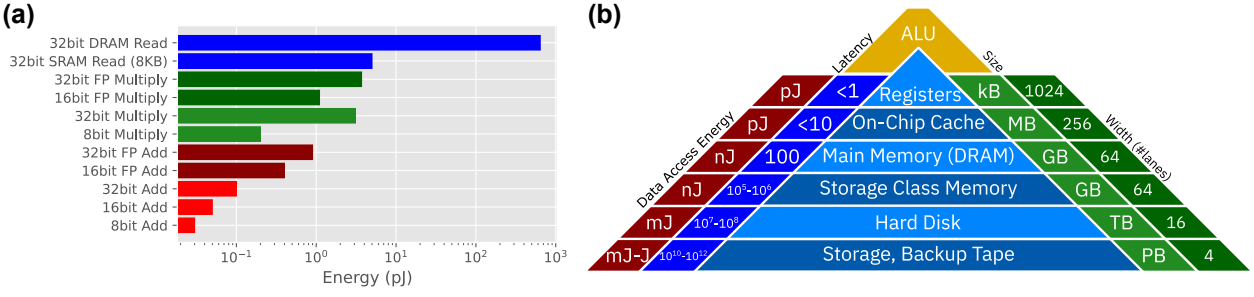

Figure 1: Typical energy required for logic operations (a) and data transfer between various levels of storage or memory and the arithmetic logic unit (ALU) (b). The impact of the logic operation accuracy and data transfer on the total power consumption is large. Data from the study by Horowitz [6].

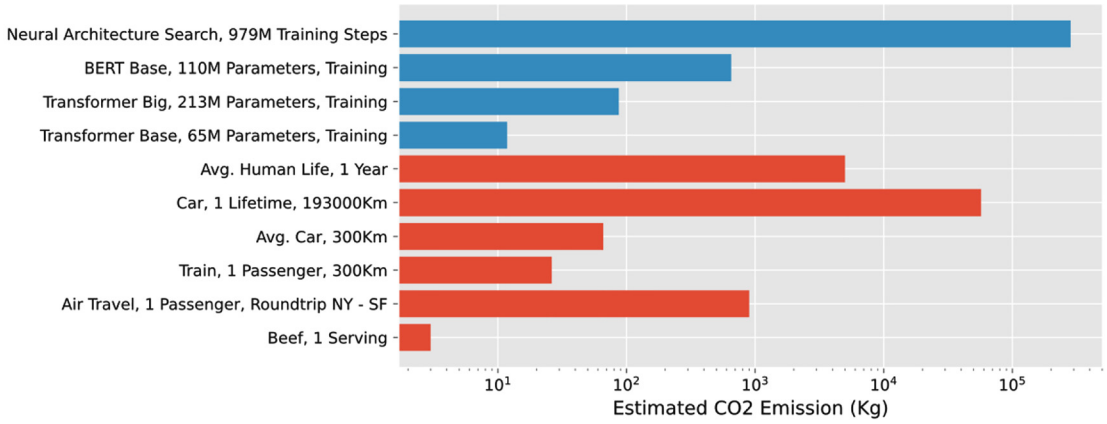

Figure 2: Comparison of the estimated, equivalent CO_2 emission for the training of different state-of-the-art deep neural networks for natural language processing (blue bars) with various everyday activities (red bars). Equivalent CO_2 for the neural networks was estimated based on the power consumption required for the training. The equivalent CO_2 emission for a neural network architecture search evolving from the Transformer big model with 979 million training steps was estimated to be 284,000 kg. Training of a single 'Transformer Big' model with 213 million parameters still emits roughly 100 kg of CO_2. Data sources: [7] for the neural network power estimations, [8] for air travel estimation and [9] for the remaining values.

In the remaining of this paper, we will address the training and inference of artificial DNNs. DNNs are the most advanced and widespread applied architectures [14]. The tremendous interest and success originates from learning algorithms based on backpropagation, with which synaptic weight values in a DNN can be optimized efficiently, such that the DNN performs a desired task [14, 15].

Fully connected feedforward neural networks with multiple hidden layers are a typical example of a DNN (Figure 3b). The basic building block of these networks is interconnected neurons, wherein each neuron applies a nonlinear activation function on the sum of all its input signals (Figure 3a). The neurons are arranged in layers, and neurons of consecutive layers are connected by a synaptic (weighted) connection. A DNN has at least two hidden layers; current state-of-the-art networks often use hundreds of hidden layers. To calculate the neural network response, a vector-matrix operation $x_i W_i$ and the nonlinear activation function are executed for each layer in the system. Evaluating the vector-matrix operation is the most compute-intensive operation in the *inference and training* of the neural network. Therefore, the synaptic interconnect will be a focus of this paper.

2 Prospects of integrated photonic neural networks

Photonic technologies are widely applied in our daily life. Integrated photonics, as in silicon photonics [16] and indium phosphide–based technologies [17], emerged in solutions for optical communication in long-range, metro and recently also short-range links. The decisive advantages of optics are the larger bandwidth-distance product, the massive parallelism, low propagation loss, density and the availability of broadband optical amplifiers. This enables the transmission of highly multiplexed signals over large distances through optical fibers [18]. A single optical amplifier restores the power of a series of wavelength-multiplexed signals, each operating at bandwidths exceeding 100 Gb/s. In such an optical link, the integrated optical technologies mentioned above provide the interface between the electrical system and the optical fiber. Important integrated optic building blocks are high-speed electro-optical modulators, detectors and a wide range of passive optical devices such as couplers, splitters and wavelength (de)multiplexers. The introduction of optical technology in data centers, for example, the large disaggregated cloud systems mentioned above, was an important step for the integrated optic technology platforms. Silicon photonics specifically profited from this new application as it provides a cost-effective scalable platform, but indium phosphide–based devices and subsystems remain vital as well. However, despite advances in integration, photonic solutions come with an overhead in terms of number of components, assembly, size, reliability and hence often cost. The penetration of integrated optic technology in computing systems showed that it must provide concrete performance or functionality advantages at reasonable cost as compared to electrical solutions to be a viable alternative [18]. Similar considerations will hold

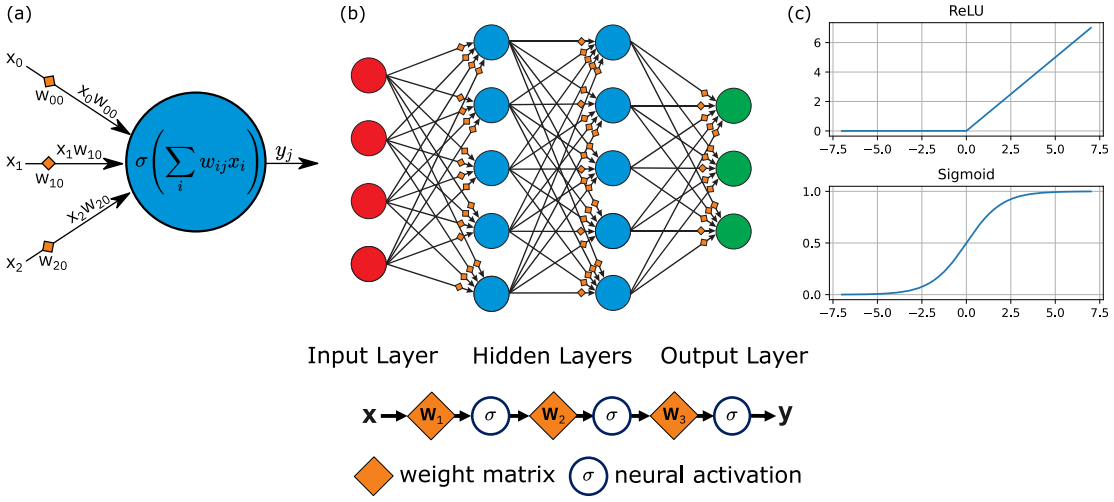

Figure 3: (a) Single neuron with synaptic connections. Each signal x_i is weighted by the corresponding synaptic connection. The weighed signals reach the neuron, where they are summed together before the nonlinear activation function σ is applied. (b) Small feedforward neural network with two hidden layers. (c) Two nonlinear activation functions that are typically used in deep neural networks (DNNs): the rectified linear unit (ReLU) and the sigmoid activation.

for photonics neuromorphic systems. Hence, carefull considerations on the specific differences optics can bring, are of utmost importance [19].

Based on the above information, optical links are a first consideration for applying photonic technology in neuromorphic computing. Similar as in large-scale computing systems, optical technology can help provide bandwidth over distances in a denser or more power-efficient way than electrical technology. However, this does not inherently change the computing architecture. Here, we focus on new applications for photonics, facilitating novel ways for performing data movement and signal processing. In the study by Denz [20] and De Marinis et al. [21], an excellent overview is given on the broad applicability of photonics for neuromorphic computing, covering optical and nonlinear photonic signal processing, materials, technology, architectures and applications. Three major advantages of photonics for neuromorphic computing are cited. We add a fourth argument:

(1) Large bandwidth, processing of high-speed data.
(2) Massive parallelism based on the 'superposition of light'.
(3) Parallel handling of images of arrays of light points – so-called pixels.
(4) The ability to process signals with low latency, real-time signal processing.

Also, in the study by Denz [20], light-matter interactions are described, providing additional functions of importance for neuromorphic computing. Some photonic materials' properties of interest are

(1) Electro-optic effect, to control the photonic signal phase by applying a current or electrical field.
(2) Electroabsorption, to impact the optical signal transmission by an electrical signal.
(3) Trimmable refractive index or absorption, which results in a persistent change in the material's optical properties by applying an optical or electrical signal. This is of special interest for nonvolatile weights.
(4) Photorefractive effect, local change of the refractive index through exposure to light.

Many nonlinear optical effects provide ultrafast response times as well as good reproducibility of the induced change of the optical properties; the electro-optic Pockels effect is an excellent example [22]. For analog signal processing, such properties are crucial as they enable fast and precise tuning of the photonic circuit functionality. Though photonic signal processing can be inherently fast, signal-to-noise considerations limit the maximum operation speed. Nevertheless, optical communication technology shows the ability to operate photonic systems in the 50-GHz or 100-Gb/s range. Though feasible, whether implementing multiplexing is a viable option depends on the problem to be solved. Wavelength division multiplexing creates an overhead that in optical communication is well justified for long-range (>10 km) but not for short-range (<150 m) links. Similar considerations will apply for photonic neuromorphic computing solutions; the option to implement an integrated multiplexed signal processing or transmission solution must be evaluated against the size, performance and cost compared to multiple single signal processing units.

A wide variety of photonics for neuromorphic computing examples is described in the literature. Though this paper focuses on integrated optics technology, it is worthwhile to also study bulk- and fiber-optic systems as examples highlighting advantages and capabilities of photonic signal processing. Parallel handling of arrays of light points, the third item in the list above, is ideally exploited in three-dimensional systems and hence a typical example of a bulk optical system. The benefits were also anticipated for optical signal processing for supercomputing [23]. An input vector is projected onto an array of spatial light modulators and subsequently on the detector array to perform, for example, an analog vector-matrix multiplication. In the study by Saade et al. [24], the spatial light modulator prepares the input vector that is imaged onto a scattering medium and from there onto a detector array. The scattering matrix technique enables a reduction in the dimension of the input data (random projection) and therefore a more efficient neuromorphic signal classification in the subsequent steps. This type of system is capable of handling high-speed optical signals, but scalability limitations are imposed by the spatial light modulator array size and update rate.

Fiber-optic systems provide an ideal means for guiding light over long distances with ultralow propagation loss. In single-mode optical fibers, the phase properties of the optical signals are preserved. This opens a path to enhanced feedback dynamics as, for example, demonstrated in the study by Brunner et al. [25]. The single-mode fiber reservoir system with a single time-modulated input signal and nonlinear optical source can demonstrate spoken digit recognition and chaotic time-series prediction at data rates beyond 1 Gb/s. Multimode fiber systems do not preserve the phase of the light, but operation is based on the signal power only. Though this may affect functionality, avoiding drift and noise of the optical phase can offer a stability advantage [26].

Integrated optic devices for neuromorphic computing offer several performance merits such as form factor, manufacturability, cost, mechanical stability and the availability of high-speed devices such as modulators and detectors. The examples above show that specific advantages exist for each implementation. The application and specific implementation decide on the viability of each approach. In the following sections, we discuss integrated optic neuromorphic computing architectures and implementations, starting with reservoir computing (RC) approaches.

2.1 Integrated photonic RC systems

RC is a computation concept well suited for sequential data processing (Figure 4) [27, 28]. A stream of input data is coupled into a reservoir, which consists of recurrently connected neurons. The synaptic interconnects between the input and the reservoir, as well as within the reservoir, are assigned randomly and kept fixed. RC systems are therefore a special type of recurrent neural networks (RNNs). The connections in the reservoir are typically sparse (<20%). To avoid exponential growth of the signals in the reservoir, the weights in the reservoir are scaled such that the system fulfills the echo state property [27]. During training, only the weights at the output layer are learned. RC has been of great interest as it massively simplifies the training compared to general RNNs. In an echo state network with a linear output layer, the weights can be learned by a simple ridge regression. While the simple training method is still beneficial, deep learning methods have made great progress over the last years and allow for very effective application of RNNs on complex tasks that could hardly be solved by RC systems. Nevertheless, RC remains an interesting concept for neuromorphic systems as the fixed reservoir weights map very well to a variety of non–von Neumann hardware implementations. Tanaka et al. [29] review various physical RC implementations ranging from electronic to optical and mechanical as well as biological implementations. Bulk, fiber and integrated *photonic* RC systems are reviewed in detail in the study by Van Der Sande et al. [30]. Here, we will give an overview of the integrated systems.

Some of the early concepts for integrated photonic reservoir systems evolved around networks of semiconductor optical amplifiers (SOAs). Each SOA provides an optical nonlinearity owing to its power saturation behavior and has a rich internal dynamic behavior. In the study by Vandoorne et al. [31], a waterfall network architecture with

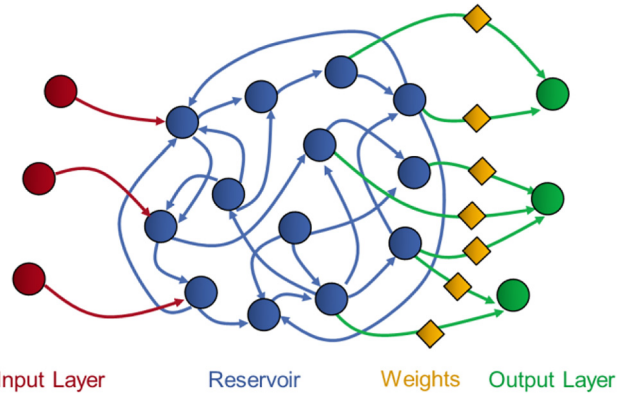

Figure 4: Illustration of the reservoir computing approach. The weights at the input and in the reservoir are randomly selected and kept fixed. The connectivity in the reservoir is sparse, and the connections are recurrent. The weights at the output layer can therefore be trained by a simple ridge regression.

feedback connections and SOA nodes was suggested. Its performance was compared to an echo state network with tanh activation function (classical software implementation) on a simple pattern recognition task, indicating a slightly better performance for the SOA network, despite the simple network architecture. Improved architectures were proposed, and the performance benefit over traditional software implementations has been demonstrated in numerical simulations for various tasks, such as, for example, spoken digit recognition [32]. However, owing to the large power consumption of the SOAs and, hence, the limited power efficiency of these networks, the first hardware realization was based on a different concept, a passive silicon photonic implementation.

The coherent silicon photonic RC implementation presented in the study by Vandoorne et al. [33] is based on a passive, linear photonic circuit. The linear reservoir nodes are arranged on a grid and connected through waveguides (delay lines), splitters and combiners to their neighbors (Figure 5a). The state at each reservoir node is read out through a detector, introducing a square nonlinearity at the output of the system; the reservoir itself remains linear. The operation speed of the reservoir is determined by the length of the waveguide delay lines between the neighboring nodes, which in this case was set to 2 cm (280 ps), to match the speed of the available measurement equipment. Using shorter delay lines, the operation speed could easily be increased to 100's of Gb/s. Operation up to 12.5 Gb/s was demonstrated for a 16-node system and based on various tasks. In the experiments, the input signal was fed into a single node, and the output signals at 11 nodes were recorded. The output signals were digitized and weighted and combined in software. The training was performed offline using ridge regression. Excellent performance was reported for timewise Boolean operations like exclusive or (XOR), bit header recognition or spoken digit recognition. Various improvements on the architecture have been applied over the years. The input scheme was optimized, by injecting the input signal to multiple nodes for a better power distribution in the network [34]. The use of multimode waveguides was suggested in the study by Katumba et al. [35] to minimize the combining loss of Y-junctions, and a novel architecture based on four-port devices for minimal loss and improved state-mixing behavior was introduced in the study by Sackesyn et al. [36] (Figure 5b). However, the missing nonlinearity inside the reservoir, the bandwidth limitations and latency imposed by detecting and weighting the output signals in the electronic domain, as well as the large number of required photodetectors for parallel operation (one detector per node), will significantly limit the practical applicability of these systems.

To overcome the later, all-optical photonic RC concepts have been suggested in the studies by Freiberger et al. [37] and Stark et al. [38] (Figure 6). The amplitude and phase of each node's output signal are weighted in the optical domain using ring resonators [39] and phase shifters or Mach-Zehnder interferometers [40], and subsequentially, the weighted signals are summed together using a coherent photonic combiner tree. The optical output signal can then either be kept in the optical domain, e.g., for nonlinear dispersion compensation in an optical link [41], or be converted into the electronic domain for further processing using a single fast photodetector. However, these systems cannot make use of the nonlinearity through the detection as the detection occurs only after the node signals have been coherently combined. It is nevertheless appealing to work with such coherent, linear photonic systems as some problems like the well-known XOR function are linearly separable in the complex domain ($\mathbb{C}$), but not in the real domain ($\mathbb{R}$) [42]. An additional challenge for all-optical systems is that the individual complex node states cannot be read out; hence, training these all-optical networks requires iterative optimization techniques to optimize both phase and amplitude of each weight. A method to reconstruct the complex states at each node is discussed in the study by Freiberger et al. [43], showing promising performance in numerical simulations. To reconstruct the amplitude of the states, all weights are set to zero, and sequentially, a single weight is enabled, and the output is recorded. In a second step, the relative phase difference between the states and a selected reference state is obtained, by turning on this pair of weights. Blackbox optimization techniques like the covariance matrix adaptation evolution strategy (CMA-ES) work as well but require a much longer training time.

One way to implement the missing nonlinearity in these systems is to embed SOAs based on III–V materials on top of the silicon photonic stack. In the study by Stark et al. [38], we demonstrated a concept for such a system (Figure 6), which is based on a four-port architecture, and besides the SOAs, uses nonvolatile optical weights based on electro-optic barium titanate technology. The main drawback of using SOAs is their large energy consumption and the rather complex fabrication process. An alternative concept to bring nonlinearity into the silicon photonic reservoir is the use of nonlinear microring resonators [44]. The nonlinearity occurs through two-photon absorption, free carrier absorption and dispersion in the ring resonator. By carefully optimizing the operation point and delay line lengths, promising performance and excellent energy efficiency was obtained in numerical simulations for a timewise XOR task.

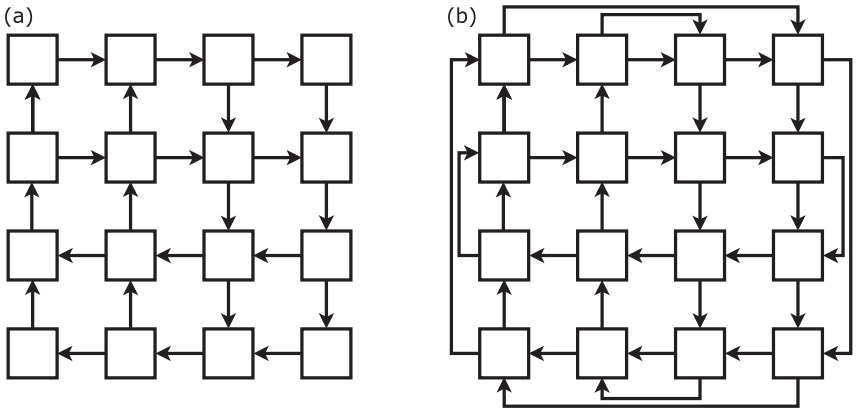

Figure 5: Two integrated photonic reservoir computing architectures were suggested in the study by Sackesyn et al. [36]. The swirl architecture (a) and the four-port architecture (b), wherein each node (black box) is connected to four other nodes. The four-port architecture offers improved power efficiency as it does not use Y-junctions but four-port devices to mix and redistribute the signals.

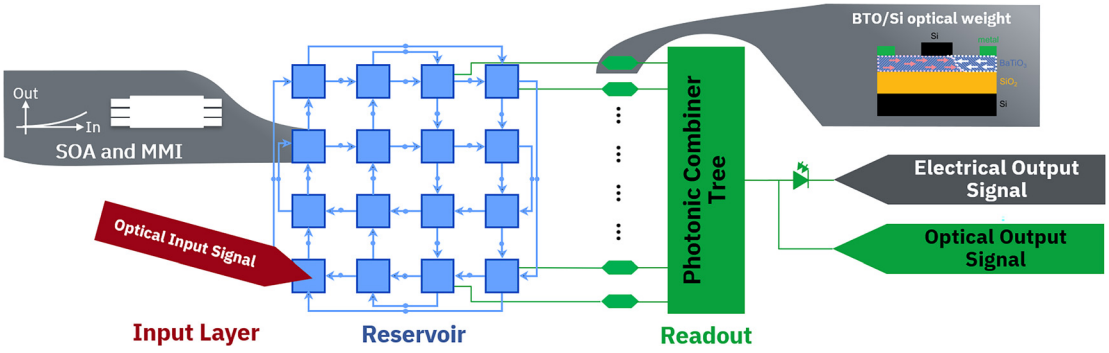

Figure 6: Concept for an all-optical integrated reservoir computing system. The reservoir consists of a network of nodes (blue boxes) based on semiconductor optical amplifiers (SOAs) and multimode interferometers (MMIs), which are connected by delay line waveguides (blue dots). For the node connection, a four-port architecture was used. Additionally, a fraction of the light is coupled out at each node and transmitted to a photonic weight. We suggest using electro-optic switches based on barium titanate to implement the signal weights. After the weighting, the signals are combined through a coherent combiner tree, and the output signal is either converted into the electrical domain using a photodetector or kept in the optical domain.

In the following sections, we will discuss two specific examples where we see potential for integrated photonic implementations of analog MAC accelerators. For comparison, we reference some bulk optical systems as well. In the first example, we focus on neural network inference, whereas the second option also opens a path toward efficient neural network training.

2.2 Integrated photonic devices for neural network inference

For neural network *inference*, there is a need for real-time power-efficient vector-matrix multiplications of high-speed signals. In literature, several examples were demonstrated of cascaded Mach-Zehnder interferometer structures performing a unitary matrix calculation. The matrix elements, representing the synaptic weights, are externally controlled by setting the phase of electro-optic tuning elements in the Mach-Zehnder arms [45]. Alternative architectures are, for example, based on ring resonator filters [46]. Barium titanate on silicon photonics electro-optic devices offer ultrahigh-speed phase modulation [22] and ultralow power tuning in the nanowatt range [47]. The maturity of state-of-the-art silicon photonics and silicon nitride and indium phosphide platforms is well suited for the implementation of this type of devices. Most demonstrations are limited to 4×4 matrix sizes, and indeed, challenges arise in scaling to larger matrices. Phase errors limit the performance, and as full control of an $N \times N$ matrix requires $2N^2$ phase shifters, setting all elements requires many electrical signals to be controlled. This makes this option well suited for small-size matrix operations ($N < 32$) as, for example, in convolutional signal processing [48]. The inference calculation is performed in a fully parallel manner, and execution time and effort do not depend on

the matrix size. An important aspect is the option to adjust the weight values very fast out of a prestored set of values. This opens the opportunity to adapt the matrix at the same speed as data is coming in, for example, to perform different types of convolutional operations. Another promising approach for neural network inference is based on the integration of phase-change materials as adjustable absorbers with integrated optic circuits [49, 50].

2.3 Integrated photonic devices for neural network training

The second exciting opportunity for integrated optic technology is related to artificial neural network *training*. Establishing an enhanced technology platform for neural network *training* is of utmost interest. Recent publications show the large environmental footprint of today's technology in neural network training, as described in the introduction [7]. There are two basic approaches to optimize the training of photonic neural networks. Either the training methods are adapted to match the system capabilities, or the operations used in a general training method like stochastic gradient with backpropagation [15] are accelerated through photonic hardware.

An example for the former concept is given in the study by Bueno et al. [51], wherein Bueno et al. implemented an iterative photonic learning method based on a greedy learning algorithm for a 4f free-space RC system with 900 nodes. A digital micromirror device is used to set the output weights, which are therefore binary. The greedy learning algorithm randomly selects an output weight and switches its state. The system performance is evaluated, and if it improved, the new output weight configuration is kept; otherwise, the previous configuration will be restored. These steps are applied iteratively, until the required performance is achieved, or the error rate converges. The training method was able to optimize the weights in about 900 learning iterations for one-step-ahead prediction of the chaotic Mackey-Glass time series, with good generalization and performance.

For the latter, we present a concept, which extends the inference calculation of the synaptic connection between two neural layers to a technology platform in which also the backpropagation and weight update steps are performed in a fully parallel manner by optical signal processing. In the Mach-Zehnder interferometer-based vector-matrix multiplication concept, the matrix element values are set by an external subsystem. Hence, changing these values in an optimization procedure would require signals to flow from the neural network output to the control system. An *in situ* training algorithm for this type of structures was proposed [52] supporting the backpropagation algorithm [53]. It is based on performing intensity measurements in the device and storing the obtained values for processing in the subsequent steps. This communication path would still introduce an information flow bottleneck and therefore limit the performance and power efficiency of the training algorithm. A local weight update mechanism is required, directly fetching the signals in the network itself. Here, we first summarize the backward propagation algorithm as this helps to understand the merits of the optical

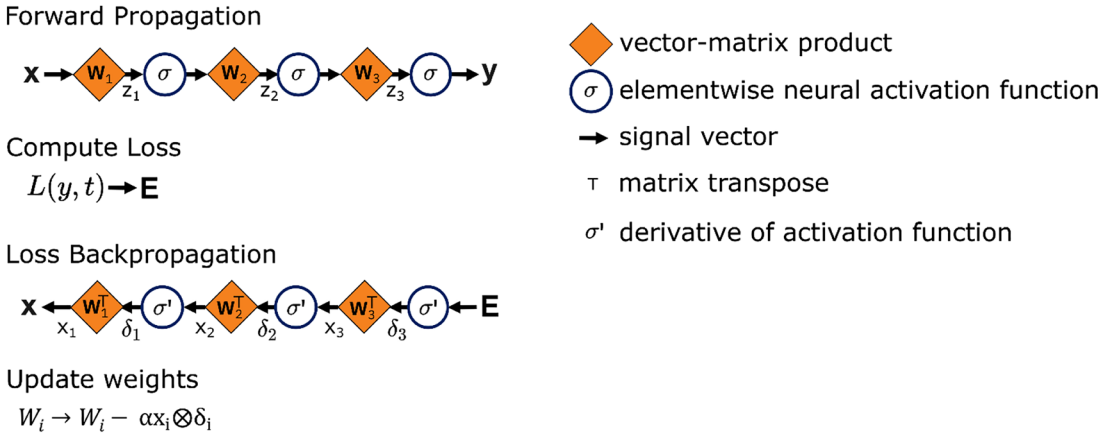

Figure 7: Illustration of forward- and backpropagation through a feedforward neural network with two hidden layers (similar to Fig. 3b) for training of the network weights. To train the network many training samples *x* with target output t are forward propagated through the network and the resulting output y are stored. Next, the loss which describes the error between the output y and the desired output is computed. Using backpropagation, we compute the error signals δ_i for each layer for all training samples. The weight updates are averaged over all training samples. This procedure is repeated iteratively until the loss is minimized.

signal processor presented thereafter. To train a feedforward DNN, we can use stochastic gradient descent together with backpropagation as follows (Figure 7) [54]:

(1) Forward propagate training input samples x_k with target response t_k and store the corresponding outputs y.
(2) For each training sample, compute the loss between the target output and the obtained outputs using a loss function. Often, the squared error is used as a loss function.
(3) For each training sample, find the error signal $\delta_i = \frac{\delta L}{\delta z_i}$, which indicates how large the influence of the input at a neuron on the total loss is. This error signals can be obtained by propagating the loss backward through the network with transposed weight matrices and using derivative of the activation functions [54]. http://neuralnetworksanddeeplearning.com/chap2.html
(4) Using the error signals obtained in step 3, update the weights as follows, to minimize the loss

$$W_i \rightarrow W_i - \alpha x_i \otimes \delta_i$$

where α is the learning rate and $x_i \otimes \delta_i = \frac{\delta L}{\delta W_i}$ is the partial derivative of the loss with respect to the weights, which is averaged over all training samples.
(5) Iteratively repeat steps 1–4 until the loss reaches a minimum.

Already in the 1990s, a photonic system was demonstrated in which the weighting elements are stored in a bulk crystal of a photorefractive material [55]. The MAC operation is obtained through the diffraction efficiency of a refractive index grating formed in the photorefractive crystal through the interference of two beams. A detailed description of signal processing in photorefractive materials is given in chapter 3 of the study by Denz [20]. In Figure 8, we depict the formation and operation principle of first a single weight and then two synaptic weights.

The diffraction efficiency of an input signal to one of the outputs represents the respective synaptic weight. Also, a part of the other input signals is diffracted toward the same output where they are coherently combined, representing the accumulation function. Because the full input vector can be applied in one inference cycle, the

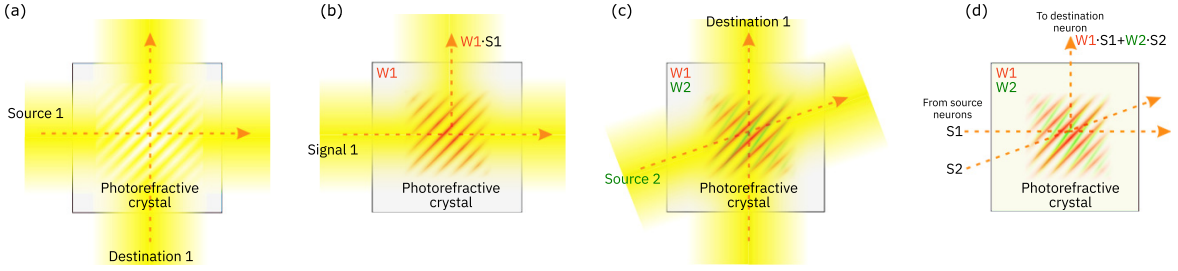

Figure 8: (a) A synaptic weight is formed in the photorefractive crystal through the interference of two light beams. Charge carriers are optically excited and diffuse to dark regions of the interference pattern. The charge separation induces an electrical field, and through the Pockels effect, a modulation of the refractive index is induced. (b) An optical signal (S1) impinging from the direction of source 1 will now be diffracted toward destination 1, resembling the operation of a single synapse. (c) A second grating is formed by applying a second source (source 2), while the destination direction is kept the same. (d) By subsequently applying input signals from the directions of source 1 and source 2, an analog multiply and accumulate operation is performed toward destination 1.

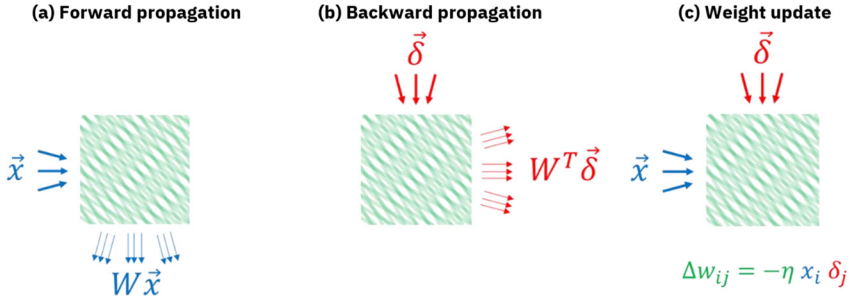

Figure 9: All critical vector-matrix calculations for neural network inference and training are efficiently performed as O(1) operations. (a) The input light is diffracted by the refractive index grating (green), which is stored in the photorefractive material. The weight matrix is given by the diffraction efficiency of the input signals to the different outputs. (b) The same refractive index grating can be used to compute the product of the transpose of the weight matrix by using alternate inputs and outputs. (c) A weight update is performed by writing a new refractive index grating by applying the input and error vector at the same time.

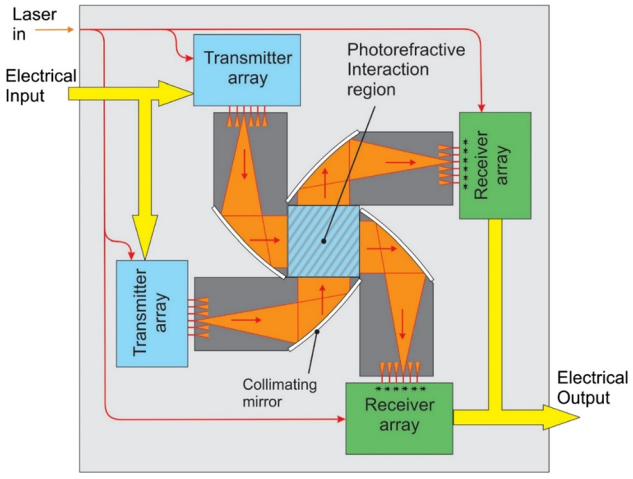

Figure 10: Schematic representation of an integrated photonic synaptic processor for inference and training. The device has one coherent optical input. The electrical input signals drive electro-optic modulators in the transmitter array sections to set the optical input vector. Collimating mirrors convert the diverging optical waves in the planar waveguide sections to collimated beams entering the photorefractive region under slightly different angles (compare with Figure 8). Both transmitter arrays are operated simultaneously to update the gratings stored in the photorefractive material. For the inference and backpropagation steps, only one transmitter array is used. The resulting optical output signal is detected by the receiver array and converted back to the electrical domain.

vector-matrix calculation is of O(1), independent of the size of the matrix. The inference calculation is now visualized for multiple beams in Figure 9a.

The backpropagation function is performed on the transpose of the matrix (W^T). The alternate input and output sections enable a straightforward calculation of the transpose operation on the same photorefractively imprinted grating, as indicated in Figure 9b. Finally, the weight update is induced by applying both the input and the error vectors at each network layer. All operations are of O(1), imposing a massive performance improvement compared to digital signal processing in von Neumann systems. Compared to the method proposed in the study by Hughes et al. [52], only intensity measurements at the device periphery are required and need to be transferred between layers of the network.

The availability of silicon photonics and the cointegration of materials like barium titanate or thin III–V layers [56] open opportunities for chip-level implementation of an analog photonic synaptic processing unit. In Figure 10, we show a device layout to implement the neural network operations based on the photorefractive effect, as depicted in Figure 9.

A thin layer of a photorefractive material is bonded to the silicon photonics wafer in which the periphery to operate the processing section is integrated. Electro-optic modulators convert the electrical input vector to the required power and phase of the optical beams. Detector arrays convert the vector-matrix output signals back to the electrical domain. To theoretically estimate the viability of this concept, we take gallium arsenide (GaAs) as the photorefractive material. The retention time of GaAs as a weight storage medium is approximately 300 ms. With a projected cycle time for the inference and backpropagation steps of 20 ns and an update cycle of 100 ns, approximately 10^4–10^6 operations can be performed before the weight values must be refreshed. Therefore, this does not represent a hurdle for the applicability of this technology for neither neural network inference nor training. The theoretical storage density is large; we evaluated that the storage of 10^6 weights in a thin layer with dimensions of 5×5 mm^2 is possible. Note that the peripheral devices must be added to the required area, resulting in a total size of 25×25 mm^2, which is still feasible. The estimated total operating power is less than 5 W. In addition to the parallelization, this in-memory computing concept avoids communication to and from memory, which is an essential aspect in meeting the power efficiency advancement. Calculations for electrical systems predict a power efficiency and performance advancement of factors larger than 100 [13]. For the optical case as presented here, we anticipate similar values as the power requirements for operating the devices are on the same order of magnitude as for the electrical memristive structure. Whether an optical implementation will be a viable solution compared to the electrical memristive structure will depend on the performance, form factor and application. Clearly, the electrical solution will have an overall larger areal density of about a factor of 20–50. Inherently, the photorefractive effect provides well-controlled setting and trimming of the weight values. This is important for efficient training and opens opportunities for analog vector-matrix multiplications with regularly updated matrix elements.

3 Conclusions

Photonic implementations of neuromorphic computing technology offer exciting properties in terms of bandwidth, processing speed and controllability. We discussed bulk, fiber-optic and integrated optic implementations of neuromorphic computing structures. The potential of integrated photonics for neural network inference and training was discussed, and a new concept for training artificial neural networks was presented. Benchmarking of results in photonic neuromorphic computing against other platforms

is important to direct the effort toward the most promising application.

Acknowledgments: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 828841 (ChipAI). This study was partially funded through the Swiss National Science Foundation grant no. 175801 Novel Architectures for Photonic Reservoir Computing.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 828841 (ChipAI). This study was partially funded through the Swiss National Science Foundation grant no. 175801 Novel Architectures for Photonic Reservoir Computing.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

[1] U. Kamath, J. Liu, and J. Whitaker, *Deep Learning for NLP and Speech Recognition*, Cham, Springer International Publishing, 2019.

[2] S. B. Furber, "Brain-inspired computing," *IET Comput. Digit. Tech.*, vol. 10, no. 6, pp. 299–305, Nov. 2016.

[3] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[4] V. K. Pallipuram, M. Bhuiyan, and M. C. Smith, "A comparative study of GPU programming models and architectures using neural networks," *J. Supercomput.*, vol. 61, no. 3, pp. 673–718, Sep. 2012.

[5] N. P. Jouppi, C. Young, N. Patil, et al., "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture*, Jun. 2017, vol. Part F1286, pp. 1–12, https://doi.org/10.1145/3079856.3080246.

[6] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb. 2014, pp. 10–14, https://doi.org/10.1109/ISSCC.2014.6757323.

[7] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," Jun. 2019 [Online]. Available at: http://arxiv.org/abs/1906.02243.

[8] A. Doherty, "Flying is bad for the planet. You can help make it better," *The* New *York Times*, 2017, https://www.nytimes.com/2017/07/27/climate/airplane-pollution-global-warming.html.

[9] "Carbon footprint," Center for Sustainable Systems, University of Michigan, 2019.

[10] G. W. Burr, R. M. Shelby, A. Sebastian, et al., "Neuromorphic computing using non-volatile memory," *Adv. Phys. X*, vol. 2, no. 1, pp. 89–124, Jan. 2017.

[11] G. W. Burr, M. J. Brightsky, A. Sebastian, et al., "Recent progress in phase-change memory technology," *IEEE J. Emerg. Sel. Top. Circuits Syst.*, vol. 6, no. 2, pp. 146–162, Jun. 2016.

[12] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nat. Electron.*, vol. 1, no. 6, pp. 333–343, Jun. 2018.

[13] T. Gokmen and Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices: design considerations," *Front. Neurosci.*, vol. 10, Jul. 2016, https://doi.org/10.3389/fnins.2016.00333.

[14] V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer, "Efficient processing of deep neural networks: a tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[16] A. E.-J. Lim, J. Song, Q. Fang, et al., "Review of silicon photonics foundry efforts," *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 4, pp. 405–416, Jul. 2014.

[17] H. Zhao, S. Pinna, F. Sang, et al., "High-power indium phosphide photonic integrated circuits," *IEEE J. Sel. Top. Quantum Electron.*, vol. 25, no. 6, pp. 1–10, Nov. 2019.

[18] M. A. Taubenblatt, "Optical interconnects for high-performance computing," *J. Light. Technol.*, vol. 30, no. 4, pp. 448–457, Feb. 2012.

[19] D. Woods and T. J. Naughton, "Photonic neural networks," *Nat. Publ. Gr.*, vol. 8, no. April, pp. 257–259, 2012.

[20] C. Denz, *Optical Neural Networks*, Braunschweig, Vieweg, 1998.

[21] L. De Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli, "Photonic neural networks: a survey," *IEEE Access*, vol. 7, no. December, pp. 175827–175841, 2019.

[22] S. Abel, F. Eltes, J. E. Ortmann, et al., "Large Pockels effect in micro- and nanostructured barium titanate integrated on silicon," *Nat. Mater.*, vol. 18, no. January, 2019, https://doi.org/10.1038/s41563-018-0208-0.

[23] H. J. Caulfield and S. Dolev, "Why future supercomputing requires optics," *Nat. Photonics*, vol. 4, no. 5, pp. 261–263, May 2010.

[24] A. Saade, F. Caltagirone, I. Carron, et al., "Random projections through multiple optical scattering: approximating Kernels at the speed of light," in *ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. – Proc.*, vol. 2016-May, pp. 6215–6219, 2016.

[25] 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.*, vol. 4, pp. 1364–1367, 2013.

[26] J. B. Heroux, H. Numata, N. Kanazawa, and D. Nakano, "Optoelectronic reservoir computing with VCSEL," in *2018 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2018, pp. 1–6, https://doi.org/10.1109/IJCNN.2018.8489757.

[27] H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Ger. Ger. Natl. Res. Cent. Inf. Technol. GMD Tech. Rep.*, vol. 148, no. 34, p. 13, 2001.

[28] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Comput. Sci. Rev.*, vol. 3, no. 3, pp. 127–149, 2009.

[29] G. Tanaka, T. Yamane, J. B. Héroux, et al., "Recent advances in physical reservoir computing: a review," *Neural Network*, vol. 115, pp. 100–123, Jul. 2019.
[30] G. Van Der Sande, D. Brunner, and M. C. Soriano, "Advances in photonic reservoir computing," *Nanophotonics*, vol. 6, no. 3, pp. 561–576, 2017.
[31] K. Vandoorne, W. Dierckx, B. Schrauwen, et al., "Toward optical signal processing using Photonic Reservoir Computing," *Opt. Express*, vol. 16, no. 15, p. 11182, Jul. 2008.
[32] K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman, "Parallel reservoir computing using optical amplifiers," *IEEE Trans. Neural Networks*, vol. 22, no. 9, pp. 1469–1481, 2011.
[33] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, et al., "Experimental demonstration of reservoir computing on a silicon photonics chip," *Nat. Commun.*, vol. 5, pp. 1–6, 2014.
[34] A. Katumba, M. Freiberger, P. Bienstman, and J. Dambre, "A multiple-input strategy to efficient integrated photonic reservoir computing," *Cognit. Comput.*, vol. 9, no. 3, 2017, https://doi.org/10.1007/s12559-017-9465-5.
[35] A. Katumba, J. Heyvaert, B. Schneider, S. Uvin, J. Dambre, and P. Bienstman, "Low-loss photonic reservoir computing with multimode photonic integrated circuits," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018.
[36] S. Sackesyn, C. Ma, A. Katumba, J. Dambre, and P. Bienstman, "A power-efficient architecture for on-chip reservoir computing," in *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions. ICANN 2019. Lecture Notes in Computer Science*, vol. 11731, I. Tetko, V. Kůrková, P. Karpov, and F. Theis, Eds., Springer, Cham., 2019, pp. 161–164.
[37] M. Freiberger, A. Katumba, P. Bienstman, and J. Dambre, "On-chip passive photonic reservoir computing with integrated optical readout," in *IEEE International Conference on Rebooting Computing (ICRC)*, Washington, DC, 2017, pp. 1–4.
[38] P. Stark, J. Geler-Kremer, F. Eltes, et al. "Novel electro-optic components for integrated photonic neural networks," in *Optical Fiber Communication Conference (OFC) 2020, OSA Technical Digest (Optical Society of America, 2020), paper M2I.4*, 2020, pp. 6–8.
[39] W. Bogaerts, P. De Heyn, T. Van Vaerenbergh, et al., "Silicon microring resonators," *Laser Photonics Rev.*, vol. 6, no. 1, pp. 47–73, 2012.
[40] P. Stark, J. Geler Kremer, F. Eltes, et al., "Non-volatile photonic weights and their impact on photonic reservoir computing systems," in *2019 Conference on Lasers and Electro-Optics Europe & European Quantum Electronics Conference (CLEO/Europe-EQEC)*, Jun. 2019, p. 1, https://doi.org/10.1109/CLEOE-EQEC.2019.8871437.
[41] A. Katumba, M. Freiberger, F. Laporte, et al., "Neuromorphic computing based on silicon photonics and reservoir computing," *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 6, pp. 1–10, 2018.
[42] I. Aizenberg, *Complex-Valued Neural Networks with Multi-Valued Neurons*, vol. 353, Berlin, Heidelberg, Springer, 2011.
[43] M. Freiberger, A. Katumba, P. Bienstman, and J. Dambre, "Training passive photonic reservoirs with integrated optical readout," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 1943–1953, 2018.
[44] F. D. Coarer, M. Sciamanna, A. Katumba, et al., "All-optical reservoir computing on a photonic chip using silicon-based ring resonators," *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 6, pp. 1–8, Nov. 2018.
[45] Y. Shen, N. C. Harris, D. Englund, and M. SoljaCiC, "Deep learning with coherent nanophotonic circuits," *Nat. Photonics*, vol. 11, no. 7, pp. 441–446, Jun. 2017.
[46] A. N. Tait, N. C. Harris, S. Skirlo, et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, p. 7430, Dec. 2017.
[47] F. Eltes, G. E. Villarreal-Garcia, D. Caimi, et al., "An integrated optical modulator operating at cryogenic temperatures," *Nat. Mater.*, Jul. 2020, https://doi.org/10.1038/s41563-020-0725-5.
[48] H. Bagherian, S. Skirlo, Y. Shen, H. Meng, V. Ceperic, and M. Soljacic, "On-chip optical convolutional neural networks," pp 1–18, 2018. [Online]. Available at: http://arxiv.org/abs/1808.03303.
[49] Z. Cheng, C. Ríos, W. H. P. Pernice, C. David Wright, and H. Bhaskaran, "On-chip photonic synapse," *Sci. Adv.*, vol. 3, no. 9, pp. 1–7, 2017.
[50] J. Feldmann, et al., "Parallel convolution processing using an integrated photonic tensor core," [Online]. Available at: https://arxiv.org/abs/2002.00281.
[51] J. Bueno, S. Maktoobi, L. Froehly, et al., "Reinforcement learning in a large scale photonic recurrent neural network," *Optica*, vol. 5, no. 6, pp. 1–5, 2017.
[52] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, vol. 5, no. 7, p. 864, Jul. 2018.
[53] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[54] M. A. Nielsen, *Neural Networks and Deep Learning*, San Francisco, CA, Determination Press, 2015.
[55] Y. Owechko and B. H. Soffer, "Holographic neurocomputer utilizing laser diode light source," in *Optical Implementation of Information Processing*, International Society for Optics and Photonics, vol. 2565, Aug. 1995, pp. 12–19.
[56] M. Seifried, G. Villares, Y. Baumgartner, et al., "Monolithically integrated CMOS-compatible III-V on silicon lasers," *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 6, 2018, https://doi.org/10.1109/JSTQE.2018.2832654.