Scalable Networks of Neuromorphic Photonic Integrated Circuits

Lei Xu, Thomas Ferreira de Lima, Hsuan-Tung Peng, Simon Bilodeau, *Graduate Student Member, IEEE*, Alexander Tait, *Member, IEEE*, Bhavin J. Shastri, and Paul R. Prucnal, *Life Fellow, IEEE*

(Invited Paper)

***Abstract*—Neuromorphic photonic integrated circuits over silicon photonic platform have recently made significant progress. Photonic neural networks with a small number of neurons have demonstrated important applications in high-bandwidth, low latency machine learning (ML) type signal processing applications. Naturally an important topic is to investigate building a large scale photonic neural networks with high flexibility and scalability to potentially support ML type applications involving high-speed processing of a high volume of data. In this paper we revisited the architecture of microring resonator (MRR) -based non-spiking and spiking photonic neurons, and photonic neural networks using broadcast-and-weight scheme. We illustrate expanded neural network topologies by cascading photonic broadcast loops, to achieve scalable neural network scalability with a fixed number of wavelengths. Furthermore, we propose the adoption of wavelength selective switch (WSS) inside the broadcasting loop for wavelength-switched photonic neural network (WS-PNN). The WS-PNN architecture will find new applications of using off-chip WSS switches to interconnect groups of photonic neurons. The interconnection of WS-PNN can achieve unprecedented scalability of photonic neural networks while supporting a versatile selection of mixture of feedforward and recurrent neural network topologies.**

***Index Terms*—Silicon photonic neural network, neuromorphic photonic computing, wavelength selective switching.**

I. Introduction

RECENT advancements in silicon photonic manufacturing have created an unprecedented opportunity to produce large-scale, low-cost photonic integrated circuits (PICs) in high volume. Neuromorphic photonic computing, inspired by the human brain's architecture of interconnected neurons that are capable of processing information much more efficiently than existing approaches, has attracted strong attention and made significant progress recently. In Fig. 1, we see the biological neuron with a model of how the signals are processed. Each individual neuron receives multiple input signals from the outputs of other neurons, applies independent weights to each of those inputs, sums the inputs, and performs a nonlinear operation to that sum before sending the output to many other neurons. This way of processing information is efficient because each operation is completed in a parallel, distributed manner. Photonic neural networks can achieve over 100x energy efficiency improvement and 1000x latency improvement in comparison with performance projections for electronic counterparts [1]. In comparison with neuromorphic electronics, neuromorphic photonic systems can operate in nanosecond time scale, or six orders of magnitude faster in some cases [2].

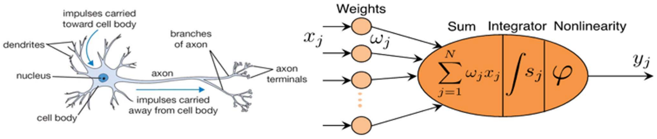

Fig. 1. Biological neuron (left) and its mathematical representation (right).

With the proliferation of optoelectronic components on photonic integration platforms, research in neuromorphic photonic computing has flourished, and machine learning and artificial intelligence (AI) algorithms running on such hardware platforms can potentially address a broad range of applications, such as medical diagnosis, telecommunications, and high-performance and scientific computing [3]. Recently, an on-chip photonic-electronic neural network comprising of a few neurons was used in a system level experiment to compensate high-bandwidth optical signal nonlinear distortions in real time after long-distance undersea transmission [4]. We expect that photonic neural networks employing a small number of neurons will be further developed, packaged, and find more applications especially in continuous time series data processing [5].

Meanwhile, the fabrication platform for silicon photonic circuits have the capability of building large scale integrated

Manuscript received 24 February 2022; revised 23 June 2022; accepted 26 September 2022. Date of publication 3 October 2022; date of current version 19 October 2022. This work was supported in part by the Office of Naval Research under Award N00014- 18-1-2297 and in part by the DARPA PEACH program under Grant AWD1006165. *(Corresponding author: Lei Xu.)*

Lei Xu, Hsuan-Tung Peng, Simon Bilodeau, and Paul R. Prucnal are with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08540 USA (e-mail: leixu@princeton.edu; hsuantungp@princeton.edu; sbilodeau@princeton.edu; prucnal@princeton.edu).

Thomas Ferreira de Lima was with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08540 USA. He is now with NEC Laboratories America Inc, Princeton, NJ 08540 USA (e-mail: thomas@nec-labs.com).

Alexander Tait is with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: atait@ieee.org).

Bhavin J. Shastri is with the Department of Physics, Engineering Physics and Astronomy, Queen's University & Vector Institute, Kingston, ON K7L 3N6, Canada (e-mail: bhavin.shastri@queensu.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSTQE.2022.3211453.

Digital Object Identifier 10.1109/JSTQE.2022.3211453

circuits. For wavelength division multiplexing (WDM) type integrated photonic neurons using microring resonator (MRR) weight banks, broadcast-and-weight architecture was proposed [6] and demonstrated [5], and multi-level broadcast-and-weight loops can be cascaded to further improve the scale of the network using a fixed and limited number of WDM wavelengths (considering less than 100 WDM wavelengths). Such architecture has been shown to be very friendly for integration [7]. When a large number of photonic neurons is interconnected to form multi-level broadcast-and-weight loops, one decision is the choice of network topology for an artificial neural network. In the current ML and AI application using neural networks, there have been different types of neural networks demonstrated most suitable for different data types and applications. To further expand the flexibility of broadcast-and-weight architecture, a wavelength selective switch (WSS) and optical coupler pair is inserted to the broadcast loop of the broadcast-and-weight architecture. We call this architecture wavelength-switched photonic neural networking (WS-PNN). Using multiple available ports, WS-PNN can be dynamically reconfigured to form different network topologies, and support a mixture of feedforward and recurrent neural networks with high flexibility. Here the WS-PNN is envisioned to be deployed with off-chip, interconnecting integrated individual broadcast-and-weight photonic neural networks.

The rest of this article is structured as follows: Section II discusses two types of photonic neurons: non-spiking type with a microring modulator and an external light source, and spiking neuron using excitable lasers. Section III presents broadcast-and-weight photonic neural networks and their topologies, including single-group and two-group photonic neurons. In Section IV, we show our novel WS-PNN and describe possible neural network reconfigurable topologies. In Section V, we discussed relevent key enabling technologies.

II. Photonic Neurons

Photonic neural networks have been demonstrated in different platforms, including free-space holographic neural network [8], [9], optical fiber based neural network [10], [11], diffractive optics [12], [13], [14], and integrated photonic circuits [15], [16], [17]. Here we focus on the WDM-compatible integrated silicon photonic neural networks using non-spiking and spiking neurons, shown in Fig. 2.

The key functions of each neuron include three parts: independent weighting of multiple inputs, summation, and nonlinear transformation. These operations can be mapped and implemented on a silicon photonic integrated circuit using MRRs as shown in Fig. 2.

A. Weighting

The multiple input signals are encoded over the intensity of lightwave at different wavelengths. Each wavelength represents the signal from a different pre-synaptic neuron. These signals are combined and coupled into a single waveguide passing through microring weight banks [18], [19]. Each of the micro-ring resonators is designed to resonate with each input wavelength and control its weight independently by fine tuning the resonance to change the transmission. The microring weight bank is implemented with in-ring N-doped photoconductive heaters [20] in the recent works [19], [21], [22]. Tait et al. [21] developed a feedback control mechanism to thermally tune the microring by adjusting the electrical current applied to the N-doped heaters. This technique has been demonstrated to perform continuous, multi-channel control with accuracy over 8 bits [22], [23], which is comparable to the resolution of matrix multipliers used in DSP ASICs.

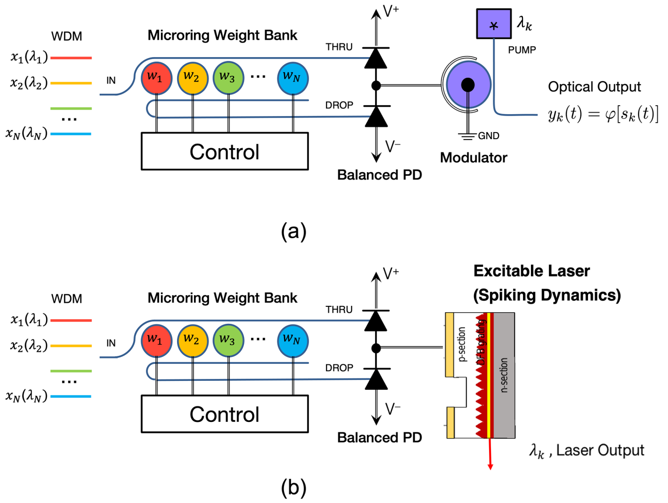

Fig. 2. Silicon photonic neuron implementation using microring resonator weight bank. (a) non-spiking neuron using an external laser source and a microring modulator and (b) spiking neurons with excitable laser.

B. Summation

The weighted optical signals after the microring weight bank are split into two output ports, corresponding to the through and drop ports of MRR, and the two output ports are optical-to-electrical (OE) converted by balanced germanium-on-silicon photodetectors (PDs) [24], which linearly transform the sum of the optical signals to photocurrent.

C. Nonlinear Transformation

The main design difference between a non-spiking and a spiking neuron is the nonliear transformation. The photocurrent is sent to a silicon micro-ring modulator (MRM) [15] as input to a non-spiking neuron as shown in Fig. 2(a), or drive a excitable laser [7] for a spiking photonic neuron as shown in Fig. 2(b). A non-spiking neuron can leverage the silicon photonics platform by integrating the weight bank, balanced photodetectors, and microring modulator on the same chip in scale, and external laser sources in different WDM wavelengths can be combined and amplified and coupled to the silicon photonic chip. In order to take advantage of silicon photonics fabrication platform, a spiking photonic neuron can be developed using both silicon photonics for microring weight banks and possible balanced photodetector, and group III-V semiconductor material for excitable lasers, with co-integration.

Due to the carrier-induced nonlinear effect, a MRM will provide the nonlinear activation to the optical pump signal. This

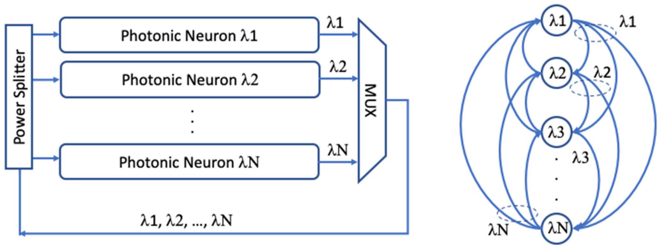

Fig. 3. Architecture of an integrated broadcast-and-weight photonic neural networks and its topology. Each of the photonic neuron has the design of either the nonspiking or spiking neurons shown in Fig. 2. Each MRR is tuned for one specific wavelength of the input signals between the on- and off- resonance states to determine the power splitting ratio between the two input ports to the balanced photodetector. MUX: wavelength division multiplexer.

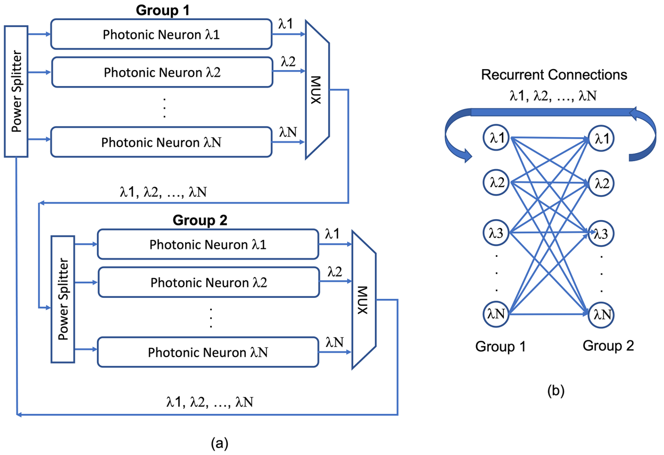

Fig. 4. (a) Broadcast-and-weight photonic neural network with two groups of photonic neurons. Each group of the neurons has the same set of output wavelengths (λ1, λ2,..., λN). (b) Topology. The group neurons have feedforward connections with a loop back for recurrent connection.

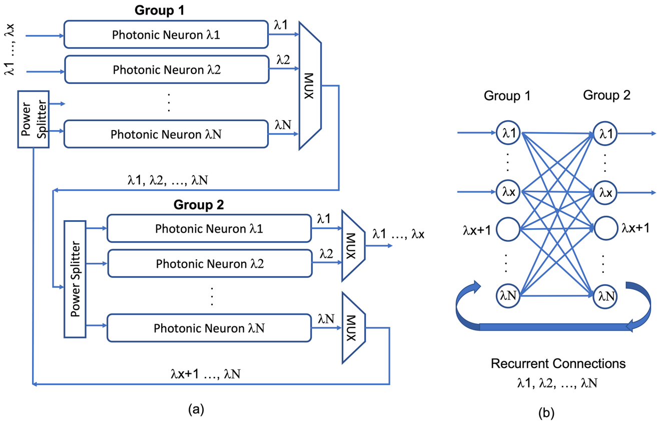

Fig. 5. (a) Broadcast-and-weight photonic neural networks with two groups of photonic neurons, and within each group, a partial number of neurons (λx+1, ..., λN) are looped back for recurrent connection. (b) Topology. The group neurons have a mixture of feedforward and recurrent connections.

mechanism shows the working principle of a neuron node on a silicon photonic circuit. The optical output of the MRM can be further sent to other photonic neuron nodes to form a network system. The microring modulator performs the nonlinear activation function. A photonic neuron will receive a combined photocurrents generated by the attached balanced photodetector, which results in the modulation of an MRM's transmission via free-carrier injection to the p-n junction. Thereby, if an extra optical source is provided and sent the input port of a MRM, the optical power will be modulated nonlinearly due to the electrical-to-optical transfer function. This nonlinear transfer function has been experimentally demonstrated to be programmable with different bias currents to the p-n junction of a MRM [15].

The distributed feedback (DFB) laser shown in Fig. 2(b) is based on a multi-quantum well (MQW) ridge-waveguide structure, electrically pumped with a p-n junction. It contains an active small section and an active large section, which are isolated by etching the p-section. The two sections are grounded to a metal pad on the chip, and each section connects to different metal pads for independent current injection. The photocurrent generated by the photodetectors flows in and out of the large section of the DFB laser, resulting in pulse-like perturbations to the laser cavity [7].

III. Broadcast-and-Weight Photonic Neural Networks

Currently, the most common photonic neural network design and implementation uses a broadcast-and-weight architecture [6], in which the input signals representing signals from other neurons are at different wavelengths, and the output is at a single wavelength. This architecture combines the outputs of multiple photonic neurons through wavelength division multiplexing and broadcast the signal to all the receiving ports of photonic neurons. This all-to-all interconnection follows a recurrent neural network (RNN) topology, as shown in Fig. 3 [18]. Incoming WDM signals are weighted by reconfigurable, continuous-valued filters called photonic weight banks and then summed by total power detection. The electrical weighted sum then modulated the corresponding WDM carrier through a nonlinear dynamical electro-optical process. Previous work on MRR weight banks have established a correspondence between weighted addition operations and integrated photonic filters. In [5], MRR weight banks were demonstrated within a broadcast-and-weight protocol and experiments.

With broadcast-and-weight architecture, Fig. 3 also shows the topological view of the neural node, and all the nodes have full recurrent connection with every nodes (including the node itself) through a single wavelength, which can be tuned continuously and independently. However, a single broadcasting loop will have limits on the number of photonic neurons, possibly due to (1) the limit number of stable wavelengths which can be deployed on a single photonic chip, (2) the optical power budget limited by the optical power splitter and the availability of optical gain, (3) the number of MRRs which can be integrated on a MRR weight bank. For a large scale photonic neural network, the broadcast-and-weight architecture proposes connecting each broadcast loop so that different loops can reuse the same set of wavelengths, similar to a cellular network in telecommunications [6], [7].

Fig. 4 and 5 show two examples of cascading broadcast-and-weight loops to form a larger photonic neural network, using two groups of photonic neural networks. For simplicity, group 1 and group 2 have the same number of photonic neurons using the same set of wavelengths (λ1, λ2,..., λN). In Fig. 4, Group 1 and Group 2 of photonic neurons are connected to form one broadcast-and-weight loop, and the photonic neurons form a recurrent-feedforward neural network (RF-NN). The outputs from the Group 1 neurons at different wavelengths (λ1, λ2,..., λN) are combined, and connected to the input optical coupler (power splitter) of the Group 2 neurons. Since the Group 2 neurons use the exact same set of wavelengths as the Group 1 neurons and pass through the same multiplexer (MUX) and power splitter, the integrated device design will be the same. Such cascading schemes can extend to multiple stages to form multi-layer feedforward neural networks, and the loop back of the output from the last group of neurons to the Group 1 input (input port of the power splitter) is optional to form recurrent neural networks. Fig. 5 shows an extension of Fig. 4, having a partial number of neurons in each group to connect with a neighboring group.

IV. Photonic Neural Networks With Wavelength Selective Switching

In section III, we reviewed the current implementation of the broadcast-and-weight architecture, showing how multiple broadcast loops reuse the same set of wavelengths to support scalable photonic neural networks. In this section, we propose scalable and flexible photonic neural networks using WSS. With WSS, a set of WDM wavelengths can be dynamically switched from one input port to multiple output ports [25]. WSS has been one of the key optical networking devices broadly used in telecommunication networks [26]. WSS can be built with micro-electromechanical systems (MEMS), liquid crystal or silicon photonics [26]. Fig. 6 and 7 show two types of WSS-based photonic neural network interconnection architecture: selected transmission and combining (STAC) model and broadcast and selection (BAS) model, respectively. Both STAC and BAS models use one WSS and one optical coupler (OC), but differentiate in having the WSS or OC at the input or output side of the photonic neural networks.

A. STAC Model

The combined outputs from the photonic neurons (λ1, λ2,..., λN) are multiplexed through a wavelength multiplexer (MUX) and sent to a WSS. Among the multiple output ports of the WSS, one of the ports (e.g. Port 1 in Fig. 6(a)) is looped back and connected to one of the input port of the optical coupler (e.g. Port 1 in Fig. 6(a)). The WSS is configured to allocate the different wavelengths to different output ports 1, 2, 3,..., M. The wavelength configuration of the WSS will decide the types of neural networks on a specific PNN chip: (1) When the WSS configures all the wavelengths to the WSS output port 1, all the channels will be looped back to the input of photonic neurons on the same chip. Therefore, a broadcast-and-weight loop is formed. (2) When the WSS configures all the wavelengths to be away from WSS output Port 1, no output from the photonic neurons from the same chip is looped back. Therefore, a single layer of neural networks is formed. (3) In general cases, the WSS can configure some of the output channels to port 1 and other channels to Port 2, 3,..., M. A partial recurrent neural network is formed through WSS Port 1 and OC Port 1 connection, and the rest of neurons are connected with other PNN chips.

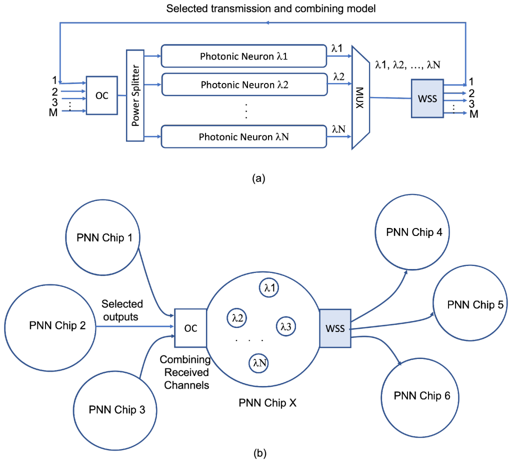

Fig. 6. (a) Photonic neural networks with wavelength selective switching (WSS): selected transmission and combining network (STAC) model. The combined wavelengths (λ1, λ2,..., λN) from the output of the photonic neurons can be switched to any output ports of the WSS, 1, 2,... M. Port 1 of the WSS is connected to the Port 1 of the optical coupler (OC) for recurrent connection. Here we set the WSS and the OC have the same number of port count, which is not necessarily needed in general network architecture of neural networks. The ports 2 to M of the WSS and OC are used for interconnecting with other photonic neural networks, which can have the same or overlapping set of wavelenghts. (b) In the selected transmission and combining network model, one photonic neural network chip (e.g. PNN Chip X) will receive selected channels from itself through WSS port 1 and other connected PNN chips (1, 2, 3 in the figure), and the OC of PNN chip X combines all the channels as the input to the photonic neurons within PNN chip X. All the WSS switches which have output ports connecting with the OC port of PNN Chip X (PNN chip 1, 2, 3 and X) need to set to avoid wavelength allocation contention.

B. BAS Model

The combined outputs from the photonic neurons (λ1, λ2,..., λN) are multiplexed through a MUX and sent to an optical coupler. The optical coupler splits the multi-wavelength optical signals to its multiple outputs, port 1, 2, 3,..., M. And each of the OC output port carries the signals from all the photonic neurons on the same chip. One of the OC output ports (e.g. Port 1 in Fig. 7(a)) is looped back and connected to the Port 1 of the WSS which is put on the input side of the photonic neural network. In the BAS model, the input ports for the WSS (Port 1, 2, 3,..., M) receive multi-wavelength signals from the photonic neural networks on the same chip and from other connected chips, as shown in Fig. 7(b). The WSS on PNN Chip X is configured to choose what sets of wavelengths from each input port to pass and combine them for the input to the photonic neural networks on this chip. The WSS can be configured similarly as in the STAC model to achieve photonic recurrent neural networks, a single layer neural network or a mixture.

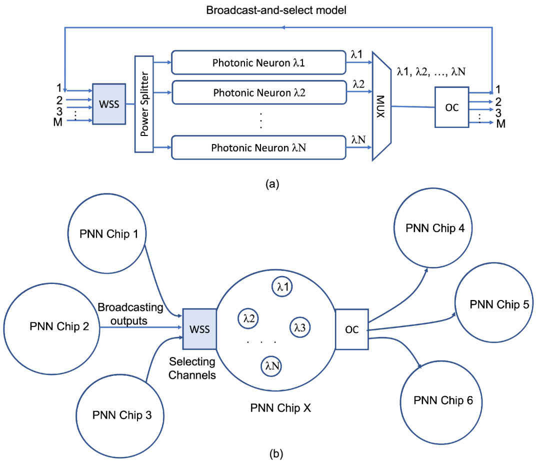

Fig. 7. (a) Photonic neural networks with WSS: broadcast and select network (BAS) model. The combined wavelengths (λ1, λ2,..., λN) from the output of the photonic neurons are broadcast to itself and other connected neurons through an optical coupler. The each of the input port (1, 2, 3,..., M) receives broadcast signals, each carrying multiple wavelengths, such as λ1, λ2,..., λN. The WSS will be configured to choose the passing wavelengths for each input port, by avoiding wavelength contention, and send the combined signals to the weight banks of each neuron. (b) In the broadcast and select network model, one photonic neural network chip (e.g. PNN Chip X) will have access to the all the output channels of the PNN Chips connected to its WSS input ports, and a single WSS on PNN Chip X can configure the wavelengths to avoid contention.

C. *Comparison of STAC and BAS Models*

As described above, the STAC model and the BAS model are able to operate the same way to configure the photonic neural networks and their interconnections. However, there exists some difference which can have important implications for network configuration and applications. (1) In the STAC model, WSS on the output side of the photonic neural network selects channels for each connected PNN chip, and the OC at the input side of the photonic neural network combines all the received channels. WSSes for different PNN chips should be configured jointly to avoid wavelength contention, which happens when two signals at the same wavelengths are combined and reach the same port. In the BAS model, a single WSS on the input side of the photonic neural network on one chip can be configured to effectively avoid wavelength contention. (2) In the BAS model, all the output signals of a photonic neural network are broadcast to a number of other photonic neural networks on different chips. While in the STAC model, only selected channels are sent to the intended photonic neural networks on different chips. When a large number of PNN chips are interconnected to form a large scale neural network to dynamically support different users, data broadcasting schemes may be regarded as being less secure and demand extra measures, such as switching on/off particular links to avoid broadcasting output signals to untended neural networks or PNN chips.

D. *Interconnected Two Wavelength Switchable Photonic Neural Networks*

The WSS-based STAC and BAS network architecture form the basic element for a WS-PNN. Considering similarity in the functions of STAC and BAS models, we will only use the STAC model to show the different WS-PNN architectures, and interconnect WS-PNN circuits to form large-scale, flexible neural networks.

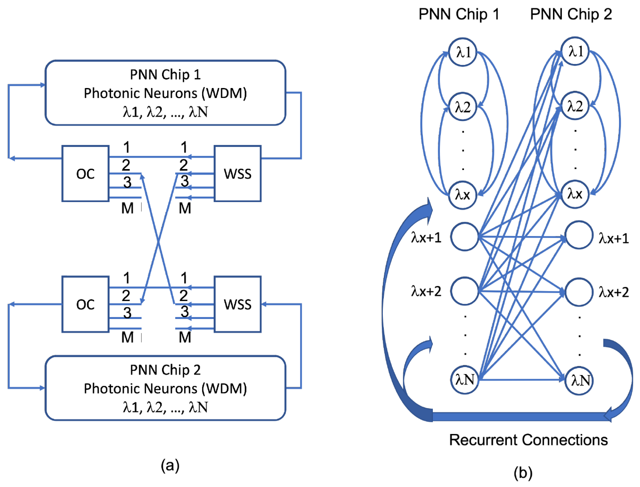

Fig. 8. (a) Two WS-PNN chips are interconnected through WSS and optical couplers. The two PNN chips are identical in having the same number of neurons with the same set of wavelengths, (λ1, λ2,..., λN). The two WSSes select the same subset of the wavelengths to interconnect the two PNN chips. (b) The architecture supports a mixture of single-layer and two-layer recurrent neural networks.

The output ports of the WSS and the input ports for the optical couplers can be used to interconnect with other WS-PNNs. In Fig. 8(a), we use two identical WS-PNNs, which have the same number of neurons on the same set of wavelengths, to be interconnected through the Port 2 of WSS and optical couplers. The two WSSes select the same set of wavelengths (λx+1, ..., λN) for interconnecting two broadcasting loops. Since each wavelength represents a specific photonic neuron within a loop, the WSS basically selects the photonic neurons (called interfacing neurons) to interconnect with neurons in other loops. In order to avoid wavelength contention, the two wavelength selective switches should have the same wavelength switching settings. In general, when multiple wavelengths which represent multiple neurons are selected to go through the Port 2 of the WSS and OC to interconnect with the other WS-PNN, the network topology shows a mixture of feedforward and recurrent neural network connections, as shown in Fig. 8(b).

To better understand the configuration of various neural network configurations using WSS, we will show the evolution of two WS-PNNs interconnection with zero, 1, multiple, all-channel interfacing wavelengths or neurons.

1) *Zero interfacing wavelength* when the WSSes on both WS-PNNs assign all the wavelengths to output Port 1, there is zero interfacing wavelength, and both WS-PNNs function as two independent broadcast-and-weight loops.

2) *One interfacing wavelength* when the WSSes on both WS-PNNs assign one wavelength (e.g. wavelength X) to output Port 2, one neuron (neuron X) from each WS-PNN is connected with the other WS-PNN, basically forming a swapping connection through one pair of photonic neurons on the two WS-PNN chips.

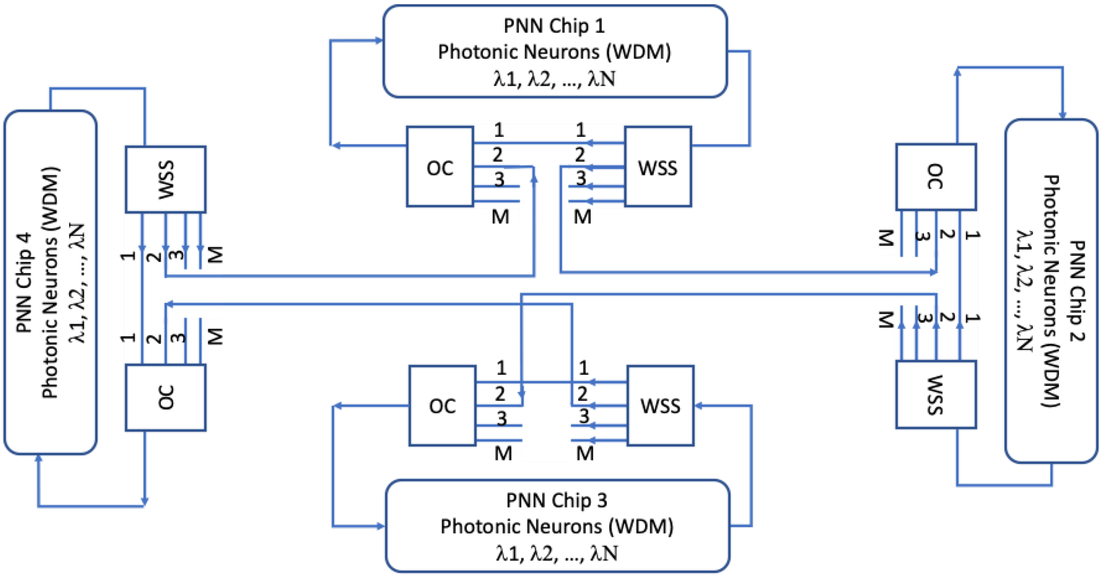

Fig. 9. Four WS-PNN chips are interconnected to form a ring. In this particular architecture, the WS-PNN chips are interconnected using the same ports and the same set of wavelengths.

For better understanding, the input signals of neuron X on PNN chip 1 are a combination of the output signals from neurons on PNN Chip 1 (excluding neuron X signal), and the wavelength X signal from PNN chip 2. Therefore, Neuron X on the two PNN chips effective carries the information from the chip it resides and sends it to the other chip it is interconned with.

3) *Multiple interfacing wavelengths* Similar to the case of one interfacing wavelength, multiple neurons from each PNN chip carrying the information from the local neural networks and the other neural networks become interfacing neurons. With multiple interfacing wavelengths, these neurons can form a multi-layer neural netwok, shown in the lower part of the neural networks in Fig. 8(b). With increasing number of wavelengths used to interconnect the two PNN chips, the neural networks can migrate to have smaller size of single-layer recurrent neural networks and larger size of two-layer recurrent neural networks.

4) *All-channel* The WSSes on the two PNN chips can configure all the wavelengths to pass through Port 2, and therefore a new two layer recurrent neural network is formed. When one of the two WSS-OC links between the two PNN chips is disconnected, a simple two layer neural network is built, and can be further connected with other PNN chips in the same way for multiple-layer neural networks.

E. Interconnected Four and Multiple Wavelength Switchable Photonic Neural Networks

The WS-PNN architecture enables a flattened neural network. Fig. 9 shows four WS-PNN chips interconnected to form a ring, in which ports 2 of PNN chips are interconnected using the same set of wavelengths, and Fig. 10 shows the topology, which shows a multi-layer feedforward and recurrent neural network architecture. The four WS-PNNs in Fig. 9 can be connected with different topology, such as a mesh, and three ports from the WSS and optical couplers will be used for interconnections, which can result different combinations of feedforward and recurrent neural networks.

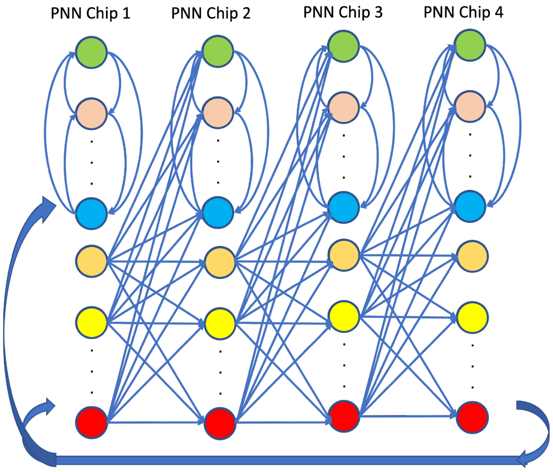

Fig. 10. The neural network topology of the photonic neurons interconnected in Fig. 9. A multi-layer feedforward neural network is formed within the topology and mixed with recurrent connections.

To further extend the scale of the WS-PNN chip interconnections, hyper cube topology can be deployed, as shown in Fig. 11. The dimension of the hypercube architecture is decided by the number of available WSS ports for interconnection. Each of the WS-PNN node will have port 1 reserved for direct connection to support recurrent connection within the WS-PNN chip itself, and the rest of the WSS and optical coupler ports can form output-input pairs to connect with other WS-PNNs. For each of the interconnection link, typically multiple wavelengths are dynamically assigned based on the intended neural network topology. In order to avoid wavelength contention, such as

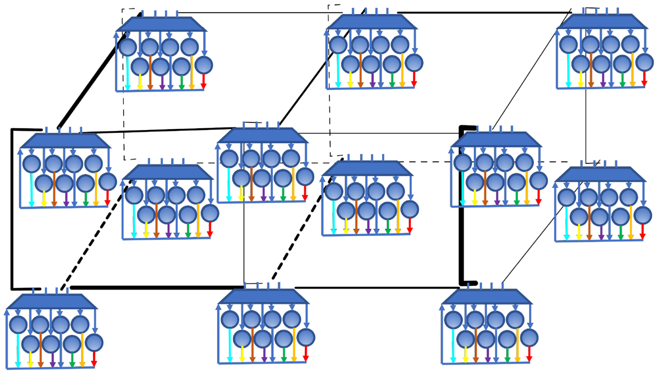

Fig. 11. Hypercube architecture formed with interconnecting the WSS and optical coupler ports of WS-PNN chips. Each of the connection can carry different numbers of wavelengths, indicated by the thickness of the link line width in the figure. The 3D cube architecture supports flattened scalability of the photonic neural networks.

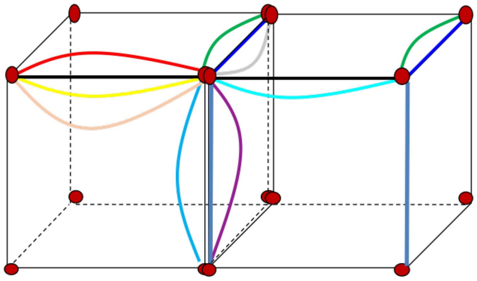

Fig. 12. Concept of edge coloring. Each node can be connected with other nodes using different wavelengths without wavelength contentions.

the same wavelength from two neighboring WS-PNN chips is assigned to the same receiving WS-PNN chip, wavelength assignment can not be arbitrary. To avoid the contention, a wavelength can only be assigned to a WS-PNN at most once. The problem becomes the edge-coloring problem on a multigraph, as shown in Fig. 12. A feasible wavelength assignment is equivalent to an assignment from the colors to the edges of the multigraph so that no two adjacent edges have the same color–exactly the edge-coloring problem. The edge coloring is a known problem, and fast heuristics are known [27] [28]. Libraries implementing this are publicly available.

V. Discussions

Neural networks are the current state-of-the-art for machine learning and there have been different topologies and layer types to choose from. Each type of neural network excels at solving a specific domain of problems, and each is tuned with hyper parameters that optimize those solutions. When building a large scale artificial neural network, the flexibility of configuring the neural networks to different topologies is important in supporting a wider range of machine learning applications. The WS-PNN architecture can support the interconnection of a large number of photonic neural networks through connecting multiple PNN chips with off-chip WSS.

A. Wavelength Selective Switches

WSS has been widely used in optical WDM communication networks to route individual wavelength for network reconfigurability. Micro ring resonator based wavelength switching architecture has been proposed for wavelength-routed optical networks-on-chip (WRONoCs) applications [29], and WRONoC leverages WDM and integrated optical circuits for multiple-processors system-on -chips [30], [31]. MRR-based WSS devices can be integrated over the same PNN chip to achieve overall compact size and scalability within the optical power and attenuation constrain before optical gain is available on silicon photonic chips. In addition, off-chip WSS devices can also be an attractive solution when considering the availability of low-cost off-chip optical gain from optical amplifiers. As an example, the free-space WSS components in [32] can have >100 WDM channels at 50 GHz channel spacing and incur insertion of about 4 dB [32]. With a k+1-ports WSS (one port is referred to the input port), one output port is reserved to connect with one of the input ports of optical coupler for the broadcasting loop within the same chip, and the rest of the k-1 ports are used for interconnection with other chips. In the hypercube architecture $k = 7$ can support flattened scaling in 6 directions (up, down, north, south, east and west). Depending on the technology platform, MEMS-based WSS can reach switching speed of a few milliseconds. A future WS-PNN system can consist of a switching matrix made of WSS and broadband optical amplifiers with a number of optical interface ports, and each of the PNN chip is packaged into line card type and plugged into the switching matrix.

B. Fault Tolerance

In a large scale photonic neural network, the key optical components can be 50 times the number of neurons, considering each neuron has a number of MRRs (equal to the number of wavelength), photodetectors, and MRM or excitable lasers. High fault tolerance to the optical component and even optical neuron failures will be critical for running processes which involve a large number of primitives. With the WS-PNN architecture, the reconfigurability of the neural networks will be at the chip level, topology level and photonic neuron level, in dealing with any failures. In [6], extra backup nodes are needed for the photonic neural nodes and the interfacing neural nodes, which increases both the hardware and management overheads. In WS-PNN architecture, interfacing neural nodes can be dynamically chosen from any of the photonic nodes within the broadcast-and-weight loop, which can effectively eliminate the needs for backup nodes and simplify the nodes management. In practice, during the WS-PNN system booting process, the management system runs self-check to identify and label failed device components and nodes, and the software algorithm will eliminate the nodes from application use.

C. Connectivity Matrix

From the perspective of neuroscience, the structure of autonomous and driven networks of spiking model neurons is

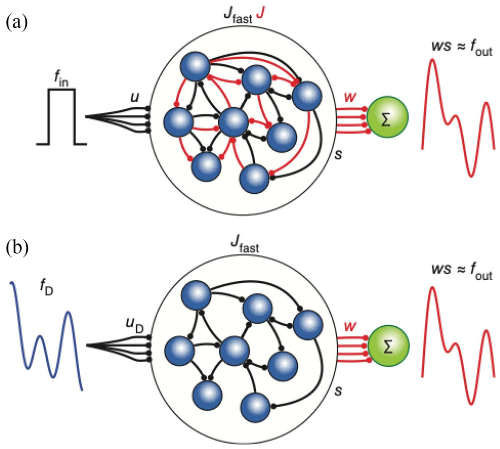

Fig. 13. Structure of (a) autonomous and (b) drive networks. The figure is taken from [33].

shown in Fig. 13 [33]. Constructing a network requires individual neuron models and synapses, a pattern of connectivity, and parameter settings. Relevant tasks defined by neuroscientists studying spiking neurons include integrating an input over time, detecting particular temporal input sequences, responding after a delay or with an activity sequence, or responding with a temporally complex output, and autonomously generating complex dynamics. Determining the connection matrix required to make a network perform a particular task is the classic credit-assignment problem of network learning. The field of machine learning has addressed credit assignment by developing error-gradient-based methods, such as back-propagation. In the case of spiking neuron networks, there are ways to solve the credit assignment problem without resorting to gradient-based procedures in [33], such as the model in [34]

A WS-PNN will generally have a mixture of feedforward and recurrent neural network. In particular for the recurrent neural network part, the recurrent time delay can have a large variance for different topologies: very small delay time (about 10 ps) for the same chip recurrent neural networks, and long delay time for cross chip recurrent neural networks when the corresponding wavelnegth goes through the optical switching matrix. Therefore, the WS-PNN is effectively a weight and delay neural network. After the intended connectivity matrix is calculated from one of the algorithms mentioned above, the WS-PNN software manager will set the topology and the weights for functional networks.

VI. Summary

With the recent progress in neuromorphic photonic integrated circuits over silicon platform, buidling a scalable network of photonic neurons going beyond tens to hundreds of them through multiple chip interconnection becomes an important research topic which will have deep impact on the future large scale neuromorphic computing systems. By including WSS in the broadcast-and-weight loop, the WS-PNN architecture brings flexibility in network topologies in addition to unprecedented scalability. Currently we limit our discussion on employing WSS off-chip to leverage commerically available WSS technologies and solutions. Meanwhile, with advancement of silicon photonic WSS technologies and solutions, WS-PNN can be integrated on the same chip for compact size. In this paper, we focus on presenting the novel concept of WS-PNN and describing the system benefits in flattened scalability and flexibility by showing selected topologies. We envision future research work around chip integration, system architecture, connectivity matrix, and applications.

References

[1] M. A. Nahmias et al., “Photonic multiply-accumulate operations for neural networks,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, pp. 1–18, Jan./Feb. 2020.

[2] T. Ferreira de Lima, B. J. Shastri, A. N. Tait, M. A. Nahmias, and P. R. Prucnal, “Progress in neuromorphic photonics,” *Nanophotonics*, vol. 6, no. 3, pp. 577–599, Mar. 2017.

[3] B. J. Shastri et al., “Photonics for artificial intelligence and neuromorphic computing,” *Nature Photon.*, vol. 15, pp. 102–114, 2021.

[4] C. Huang et al., “Silicon photonic-electronic neural network for fibre nonlinearity compensation,” *Nature Electron.*, vol. 4, no. 11, pp. 837–844, 2021.

[5] A. N. Tait et al., “Neuromorphic photonic networks using silicon photonic weight banks,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017.

[6] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Broadcast and weight: An integrated network for scalable photonic spike processing,” *J. Lightw. Technol.*, vol. 32, no. 21, pp. 3427–3439, 2014.

[7] H.-T. Peng, M. A. Nahmias, T. F. De Lima, A. N. Tait, and B. J. Shastri, “Neuromorphic photonic integrated circuits,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, pp. 1–15, Nov./Dec. 2018.

[8] D. Psaltis, D. Brady, and K. Wagner, “Adaptive optical networks using photorefractive crystals,” *Appl. Opt.*, vol. 27, no. 9, pp. 1752–1759, May 1988.

[9] D. Psaltis et al., “Optoelectronic implementations of neural networks,” *IEEE Commun. Mag.*, vol. 27, no. 11, pp. 37–40, Nov. 1989.

[10] B. J. Shastri et al., “Spike processing with a graphene excitable laser,” *Sci. Rep.*, vol. 6, 2016, Art. no. 19126.

[11] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, “Parallel photonic information processing at gigabyte per second data rates using transient states,” *Nature Commun.*, vol. 4, 2013, Art. no. 1364, doi: 10.1038/ncomms2368.

[12] X. Lin et al., “All-optical machine learning using diffractive deep neural networks,” *Science*, vol. 361, no. 6406, pp. 1004–1008, 2018.

[13] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, “Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification,” *Sci. Rep.*, vol. 8, 2018, Art. no. 12324.

[14] T. Zhou et al., “In situ optical backpropagation training of diffractive optical neural networks,” *Photon. Res.*, vol. 8, pp. 940–953, 2020.

[15] A. N. Tait et al., “Silicon photonic modulator neuron,” *Phys. Rev. Appl.*, vol. 11, no. 6, 2019, Art. no. 064043.

[16] Y. Shen et al., “Deep learning with coherent nanophotonic circuits,” *Nature Photon.*, vol. 11, no. 7, pp. 441–446, 2017.

[17] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. Pernice, “All-optical spiking neurosynaptic networks with self-learning capabilities,” *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.

[18] A. N. Tait et al., “Microring weight banks,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 22, no. 6, pp. 312–325, Nov./Dec. 2016.

[19] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Multi-channel control for microring weight banks,” *Opt. Exp.*, vol. 24, no. 8, pp. 8895–8906, Apr. 2016.

[20] H. Jayatilleka et al., “Crosstalk in soi microring resonator-based filters,” *J. Lightw. Technol.*, vol. 34, no. 12, pp. 2886–2896, Jun. 2016, doi: 10.1109/JLT.2015.2480101.

[21] A. N. Tait et al., “Feedback control for microring weight banks,” *Opt. Exp.*, vol. 26, no. 20, pp. 26422–26443, 2018.

[22] C. Huang et al., “Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits,” *APL Photon.*, vol. 5, no. 4, 2020, Art. no. 0 40803.

[23] W. Zhang et al., “Silicon microring synapses enable photonic deep learning beyond 9-bit precision,” *Optica*, vol. 9, no. 5, pp. 579–584, May 2022.

[24] M. S. Hai, M. N. Sakib, and O. Liboiron-Ladouceur, “A 16 GHz silicon-based monolithic balanced photodetector with on-chip capacitors for 25 Gbaud front-end receivers,” *Opt. Exp.*, vol. 21, no. 26, pp. 32680–32689, 2013.

[25] J. Tsai and M. C. Wu, “A high port-count wavelength-selective switch using a large scan-angle, high fill-factor, two-axis MEMS scanner array,” *IEEE Photon. Technol. Lett.*, vol. 18, no. 13, pp. 1439–1441, Jul. 2006.

[26] M. C. Wu, O. Solgaard, and J. E. Ford, “Optical MEMS for lightwave communication,” *J. Lightw. Technol.*, vol. 24, no. 12, pp. 4433–4454, Dec. 2006.

[27] J. Misra and G. Gries, “A constructive proof of Vizing’s theorem,” *Inf. Process. Lett.*, vol. 41, no. 3, pp. 131–133, 1992.

[28] I. Duarte, L. Cancela1, and J. Rebola, “Graph coloring heuristics for optical networks planning,” in *Proc. Telecoms Conf.,* 2021, pp. 1–6.

[29] Z. Zheng, M. Li, T.-M. Tseng, and U. Schlichtmann, “Light: A scalable and efficient wavelength-routed optical networks-on-chip topology,” in *Proc. 26th Asia South Pacific Des. Automat. Conf.*, 2021, pp. 568–573.

[30] M. Tala, M. Castellari, M. Balboni, and D. Bertozzi, “Populating and exploring the design space of wavelength-routed optical network-on-chip topologies by leveraging the add-drop filtering primitive,” in *Proc. 10th IEEE/ACM Int. Symp. Netw.-on-Chip*, 2016, pp. 1–8.

[31] M. Li, T.-M. Tseng, D. Bertozzi, M. Tala, and U. Schlichtmann, “CustomTopo: A topology generation method for application-specific wavelength-routed optical NoCs,” in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2018, pp. 1–8.

[32] D. M. Marom et al., “Wavelength-selective 1xK switches using free-space optics and MEMS micromirrors: Theory, design and implementation,” *J. Lightw. Technol.*, vol. 23, no. 4, pp. 1620–1630, Apr. 2005.

[33] L. F. Abbott, B. DePasquale, and R.-M. Memmesheimer, “Building functaional networks of spiking model neurons,” *Nature Neurosci.*, vol. 19, pp. 350–355, 2016.

[34] C. Eliasmith, “A unified approach to building and controlling spiking attractor networks,” *Neural Comput.*, vol. 17, no. 6, pp. 1276–1314, Jun. 2005.

Lei Xu received the B.S. degree in geophysics from Peking University, Beijing, China, in 1997, the M.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2000, and the Ph.D. degree from Princeton University, Princeton, NJ, USA, in 2005. He is currently a Research Scholar with the Lightwave Lab, Electrical and Computer Engineering Department, Princeton University. He was a Senior Research Staff Member at NEC Labs America from 2005 to 2012, and a technology co-founder of Sodero Networks, Torray Networks, and Eagle Nebula Inc from 2012 to 2019. He has authored or coauthored more than 100 research papers, and has 59 U.S. patents. His research interests include neuromorphic photonic computing, silicon photonics, software defined optical networking, and high-speed optical communications.

Thomas Ferreira de Lima received the B.Sc. and Ingénieur Polytechnicien Master degrees in 2016 from Ecole Polytechnique, Palaiseau, France, with a focus on physics for optics and nanosciences, and the Ph.D. degree in electrical engineering from Lightwave Communications Research Laboratory, Department of Electrical Engineering, Princeton University, Princeton, NJ, USA, in 2022. He is currently a Researcher with the NEC Laboratories America, Inc., Princeton. He is also a contributing author to the textbook, *Neuromorphic Photonics*. His research interests include integrated photonic systems, nonlinear signal processing with photonic devices, spike-timing-based processing, ultrafast cognitive computing, and dynamical light–matter neuro-inspired learning and computing.

Hsuan-Tung Peng received the B.S. degree in physics from National Taiwan University, Taipei, Taiwan, in 2015, and the M.A. degree in electrical engineering in 2018 from Princeton University, Princeton, NJ, USA, where he is currently working toward the Ph.D. degree. His research interests include neuromorphic photonics, photonic integrated circuits, and optical signal processing.

Simon Bilodeau (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in physics from McGill University, Montreal, QC, Canada. He is currently working toward the Ph.D. degree in electrical and computer engineering with Lightwave Communications Research Lab, Princeton University, Princeton, NJ, USA. He conducted nanoelectronics research at ultralow temperatures with McGill University. His research interests include nanophotonics, optoelectronics, integrated photonic systems, and neuromorphic computing.

Alexander Tait received the Ph.D. degree from the Lightwave Communications Research Laboratory, Department of Electrical Engineering, Princeton University, Princeton, NJ, USA under the direction of Paul Prucnal. He is currently an Assistant Professor of electrical and computer engineering with Queen’s University, Kingston, ON, Canada. He was a NRC Postdoctoral Fellow with Quantum Nanophotonics and Faint Photonics Group, National Institute of Standards and Technology, Boulder, CO, USA. He has authored 15 refereed journal papers, (co) filed eight provisional patents, created seven open-source software packages, and contributed to the textbook *Neuromorphic Photonics*. His research interests include silicon photonics, neuromorphic engineering, and superconducting optoelectronics. Dr. Tait was the recipient of the National Science Foundation (NSF) Graduate Research Fellowship (GRFP) and is a Member of the IEEE Photonics Society and the Optical Society of America (OSA). He was also the recipient of the Award for Excellence from the Princeton School of Engineering and Applied Science, the Best Student Paper Award from the 2016 IEEE Summer Topicals Meeting Series, and the Class of 1883 Writing Prize from the Princeton Department of English.

Bhavin J. Shastri (Member, IEEE) is currently an Assistant Professor of engineering physics at Queen’s University, Kingston, ON, Canada, and a Faculty Affiliate with the Vector Institute for Artificial Intelligence, Canada. He earned the B.Eng. (with distinction, Hons.), M.Eng., and Ph.D. degrees in electrical engineering (photonics) from McGill University, Montreal, QC, Canada, in 2005, 2007, and 2012, respectively. He was an NSERC and Banting Postdoctoral Fellow (2012–2016) and an Associate Research Scholar (2016–2018) with Princeton University, Princeton, NJ, USA. He has authored or coauthored more than 70 journal articles and 90 conference proceedings, seven book chapters, and given more than 60 invited talks and lectures including five keynotes and three tutorials. He is a co-author of the book, *Neuromorphic Photonics* (Taylor & Francis, CRC Press, 2017). His research interests include silicon photonics, photonic integrated circuits, neuromorphic computing, and machine learning.

Dr. Shastri is the winner of the 2020 IUPAP Young Scientist Prize in Optics for his pioneering contributions to neuromorphic photonics from the ICO. He is a Senior Member of the OSA. He was the recipient of the 2014 Banting Postdoctoral Fellowship from the Government of Canada, the 2012 D. W. Ambridge Prize for the top graduating Ph.D. student at McGill, an IEEE Photonics Society 2011 Graduate Student Fellowship, a 2011 NSERC Postdoctoral Fellowship, a 2011 SPIE Scholarship in Optics and Photonics, a 2008 NSERC Alexander Graham Bell Canada Graduate Scholarship, including the Best Student Paper Awards at the 2014 IEEE Photonics Conference, 2010 IEEE Midwest Symposium on Circuits and Systems, the 2004 IEEE Computer Society Lance Stafford Larson Outstanding Student Award, and the 2003 IEEE Canada Life Member Award.

Paul R. Prucnal (Life Fellow, IEEE) received the A.B. degree *(summa cum laude)* in mathematics and physics from Bowdoin College, and the M.S., M.Phil., and Ph.D. degrees in electrical engineering from Columbia University, New York, NY, USA. After his doctorate, Prucnal joined the faculty, Columbia University, where, as a Member of the Columbia Radiation Laboratory, he performed groundbreaking work in OCDMA and self-routed photonic switching. In 1988, he joined the Faculty with Princeton University, Princeton, NJ, USA. His research on optical CDMA initiated a new research field in which more than 1000 papers have since been published, exploring applications ranging from information security to communication speed and bandwidth. In 1993, he invented the Terahertz Optical Asymmetric Demultiplexer, the first optical switch capable of processing terabit per second (Tb/s) pulse trains. Prucnal is author of the book, *Neuromorphic Photonics*, and Editor of the book, *Optical Code Division Multiple Access: Fundamentals and Applications*. He was an Area Editor of IEEE Transactions on Communications. He has authored or coauthored more than 350 journal articles and book chapters, and holds 28 U.S. patents. He is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), the Optical Society of America (OSA) and the National Academy of Inventors (NAI), and a Member of honor societies including Phi Beta Kappa and Sigma Xi. He was the recipient of the 1990 Rudolf Kingslake Medal for his paper entitled Self-routing photonic switching with optically-processed control, received the Gold Medal from the Faculty of Mathematics, Physics and Informatics at the Comenius University, for leadership in the field of Optics 2006 and has won multiple teaching awards at Princeton, including the E-Council Lifetime Achievement Award for Excellence in Teaching, the School of Engineering and Applied Science Distinguished Teacher Award, The President’s Award for Distinguished Teaching. He has been instrumental in founding the field of Neuromorphic Photonics and developing the photonic neuron, a high speed optical computing device modeled on neural networks, and also integrated optical circuits to improve wireless signal quality by cancelling radio interference.