16 Neuromorphic Silicon Photonics

*S. Bilodeau, T. Ferreira de Lima, C. Huang,
B. J. Shastri, and P. R. Prucnal*

CONTENTS

16.1 INTRODUCTION

Neural networks have enjoyed renewed popularity over the last decade under the appellation of "deep learning" [1,2]. The idea of mimicking the brain to process information, however, can be traced back half a century prior to Rosenblatt's perceptron [3], and the first experimental models of biological neurons to Hodgkin and Huxley a few years prior [4]. The artificial neurons that make up neural networks take many forms, some more closely related to this biological inspiration. Yet all neural networks take the form of simple nodes that (a) perform a linear operation on multiple other neurons' outputs, (b) integrate the resulting signals, and (c) perform a nonlinear transformation on the summed, weighted inputs. Various interconnection topologies—feedforward, feedback (recurrent), close-neighbor translationally-invariant (convolutional), etc.—endow the network with different computational properties.

Such an asynchronous, parallel framework is at odds with the digital von Neumann architecture that electronic microprocessors often employ for their emulation. This mismatch was recognized early on, leading to pioneering work by VLSI engineers starting in the 1980s to map the physics of transistors to neuronal models for gains in computational density, energy efficiency, and speed [5]. However, Moore's Law and Dennard scaling kept such "neuromorphic" architecture outside of the limelight in favor of general-purpose digital processors. Today, this scaling nears its end, and researchers turn to ever more specialized hardware such as graphical processing units [6], tensor processing units [7], and specially-configured field-programmable gate arrays [8] to run demanding neural network models. This is renewing interest in neuromorphic application-specific integrated circuits (ASICs), the extrapolated conclusion of this trend.

Since the requirements of neuromorphic hardware differ from von Neumann digital computing, it is not obvious that silicon microelectronics must provide the best substrate for neuromorphic ASICs [9,10]. The reliance of neural networks on simple networked nodes suggests that a platform suited for communications, such as photonics, might have an advantage. This was recognized in the 80s [11], yet the lack of integrability limited investigations at the time. The commercial silicon

DOI: 10.1201/9780429292033-16

photonic platforms that have arisen over the last few years, however, now offer high index contrast, low-loss waveguides integrated with high bandwidth optoelectronics for signal modulation and detection [12]. Furthermore, the reuse of materials and processes from microelectronics allows the platform to enjoy its economies of scale. This, combined with the intrinsic appeal of photonics to emulate neural models, is one of the reasons that the newly termed field of *neuromorphic photonics* has attracted considerable attention [10,13–17].

In this chapter, we summarize past work and outline future directions of neuromorphic photonics, with a focus on silicon photonics implementations. We take a hierarchical approach. First, suitable neuronal models and the instantiation of their components is reviewed. Networking techniques to achieve neural networks proper is then discussed. Next, we review applications of neuromorphic photonics. We finish with a short outlook.

16.2 SILICON PHOTONIC NEURONS

A major impetus of the resurgence of neuromorphic photonics in the 2010s was the recognition that the dynamics of some active photonic components are mathematically equivalent to leaky integrate-and-fire spiking neuron models [18]. Silicon's indirect bandgap precludes efficient light sources and amplifiers that would easily allow such a quantum-level spiking neuron model to be implemented. This can be overcome by, for example, depositing optically active films like phase-change materials [19] or by combining emission from emissive centers with single-photon detectors at cryogenic temperatures [20].

There are, however, alternate and easier to program neuron models that lend themselves almost perfectly to the room-temperature high-bandwidth optoelectronics of silicon photonic. Such continuous artificial neurons encode information in an analog property of the light instead of spike timing:

$$\mathrm{ds(t)/dt} = \mathbf{W}\cdot \mathrm{f}(\mathbf{s}(\mathrm{t})) - \mathrm{s(t)}/\tau + \mathrm{w_{in}u(t)}. \tag{16.1}$$

Here, a neuron's internal state s drives others through a continuous (non-spiking) nonlinear transfer function. This model has two important components: (1) matrix-vector multiplication $\mathbf{W} \cdot \mathbf{x}$ (equivalently multiply-accumulate (MAC) operations) between a neuron's inputs x and its weights w, the non-biological equivalent of a synapse, and (2) a nonlinear transformation of the input state to the broadcast output $f(s)$. τ captures the time constant of the nonlinear unit, and u is the external drive. Stripped of temporal components, this reduces to the non-dynamical artificial neuron model ubiquitous in deep learning for the neuron's output $f(s) = f(\mathbf{W} \cdot f(\mathbf{s}) + w_{in}u)$. This artificial neural network model, while simple, has been immensely successful in applications and is almost exclusively used. Implemented in silicon photonics, it offers a way to do complex neural computation with nanosecond latencies, opening up a wealth of new application domains.

16.2.1 Multiply-Accumulate Operation

Two broad philosophies have been explored for multiply-accumulate operations in silicon photonics: coherent and incoherent. In the coherent framework pictured in Figure 16.1a-b, beamsplitters and phase shifters control the interference of light of a well-defined wavelength, mode, and polarization. When meshed appropriately, any unitary transformation can be performed on a path-encoded coherent input beam, directly implementing matrix-vector multiplication at the speed of light in the waveguide. This approach was originally considered for linear photonic quantum information processing, and was demonstrated with a mesh of Mach-Zehnder modulators [22]. Since the coherent approach is isomorphic to a vector-matrix multiplication, summing occurs naturally.

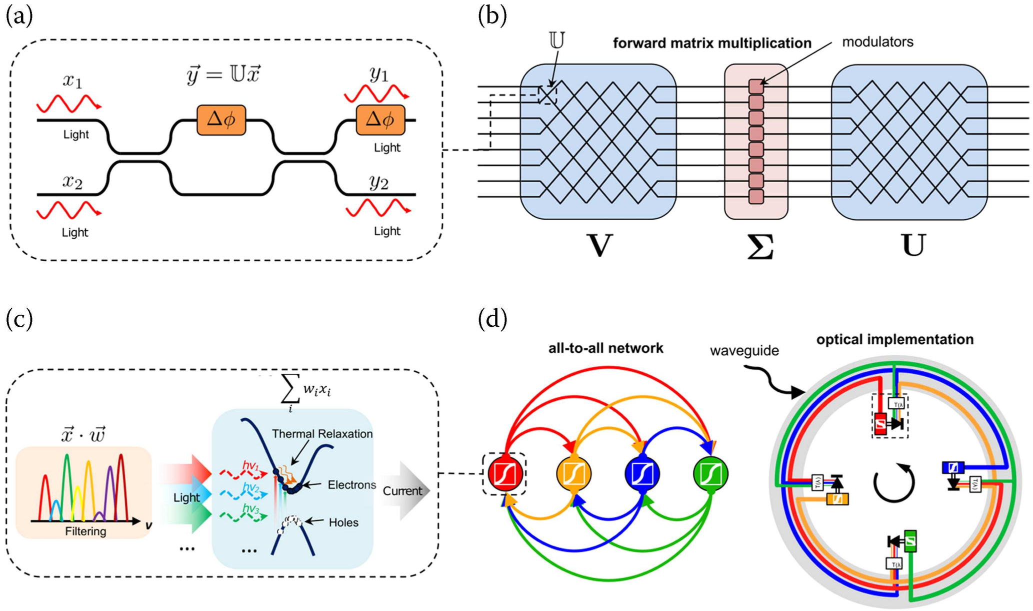

FIGURE 16.1 Coherent (top) and incoherent (bottom) multiply-accumulate operations. (a) Coherent approaches can apply unitary rotation to incoming lightwaves. This unit can perform a tunable 2 × 2 unitary rotation denoted by U. (b) Example of scaling the system to perform a matrix operation in a feedforward topology, using a U unit at each crossing together with singular value decomposition. (c) Incoherent approaches can directly perform dot products on optically multiplexed signals. However, they rely on detectors and O/E conversion for summation. (d) The ability to multiplex allows for network flexibility, which can enable larger-scale networks with minimal waveguide usage. Figure and caption adapted from [21], with permission

The incoherent approach, in contrast, is presented in Figure 16.1c-d. There, values are represented through the relative intensities of light in a collection of wavelengths, modes and/or polarization. Linear operations are performed through selective filtering and/or attenuation. In this approach, a single photodetection step implements summing by yielding a photocurrent proportional to the sum of the optical powers across the incoming modes. Termed “broadcast-and-weight”, this is explored in [23] and first demonstrated in silicon by controlling the transmission of microring resonators in [24]. Resonators can also only be used for filtering and followed by electro-absorption modulators to achieve the same effect [25].

To be able to adjust a neuron’s weights, the beamsplitters, phase shifters, filters, and attenuators mentioned above must be tunable. This constitutes the main source of complexity in a photonic neural network, since e.g., a fully connected network of N neurons will require N^2 weights to be controlled. Silicon exhibits a strong thermo-optic effect, and local metal or doped heaters are often used for “slow” index changes to implement this reconfigurability. Phase-change materials have also been considered for non-volatile control of transmission [26,27]. A technique worth mentioning and used for incoherent networks is resonator photoconductive control. First applied to silicon photonic neuron weights in [28], this technique leverages the measurable photoabsorption-induced change of a doped ring resonator’s resistance to “lock” the filter transmission to the desired point. Using photoresistance as a proxy for optical power further has the advantage of not requiring access to the optical signal for calibration, enabling large-scale actuation [29]. Control of this transmission is what ultimately limits the effective fixed-point bit resolution that can be achieved for the multiply-accumulate operation, an important metric for comparison with digital alternatives. Currently, record 7.2 bits of accuracy and precision were demonstrated [29], close to the 8-bit reduced precision popular in digital deep learning, and above some electronic neuromorphic architectures (for instance, IBM TrueNorth's 4 + 1 (sign) bits [30]). In practice, controlling and reconfiguring these weights will require a full-scale neuromorphic photonic processing system. This will include the silicon photonic chip itself, copackaged with requisite laser sources, microcontroller, and RF interfaces [31].

The performance of passive photonics for MAC operations was compared to electronics in [21]. In terms of limits of analog compute, photonics shares a similar implementation strategy as resistive (or memristive) crossbar arrays, with one part of the MAC held fixed (weight) and the other fast changing (input). This leads to similar fundamental limits. For state-of-the-art crossbar and photonic components, on-chip aJ/MAC efficiencies and 100s of PMACs/s/mm compute densities are possible. A photonic core scales better in terms of energy per MAC and compute density for (1) >100 μm core sizes, since crossbars see reduced bandwidth with length unlike waveguides, (2) for >500 channels due to the $O(N)$ scaling of photodetector capacitance compared to $O(N^2)$ crossbar capacitance scaling, and (3) for low (<4 bits) of fixed-point resolution, due to photonics having extra (shot) noise increasing with power. Future such analyses should account for all other power consumption (control electronics, memory access, lasers, and E/O, O/E conversions if required). In any case, the end-to-end latency of passive photonic MACs can be lower than electronics, leading to unique application areas that will be explored in Section 16.3.

16.2.2 Nonlinear Transformation

The core of the neuron is its internal dynamics, leading to the nonlinear transformation it performs on its weighted, summed inputs. For path-encoded coherent beams, in theory any optical nonlinearity could be used to perform an all-optical nonlinear transformation. Materials deposited on waveguides such as phase-change [32] have demonstrated such functionality. They, however, lack reconfigurability once fabricated. This can be remedied with tunable silicon photonic devices at the cost of footprint [33,34]. The all-optical nonlinear approaches above require high optical powers, however, and so local opto-electronic conversions with a tap and detector that self-modulates phase were demonstrated [35]. The incoherent approach, on the other hand, already relies on photodetection for summing. The photocurrent can be used to actuate a wavelength (or mode, or polarization)-selective amplitude modulator, whose transfer function implements an effective optical-optical nonlinearity. In [36], this is achieved with a microring modulator driven by a photodetector output and is reproduced in Figure 16.2. While the full microring transfer function is

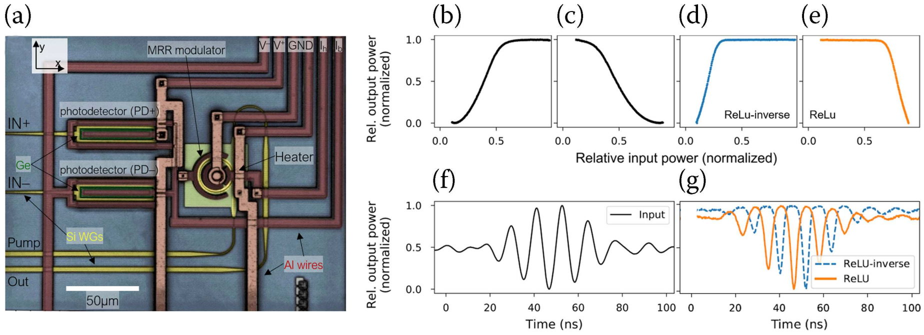

FIGURE 16.2 Example of a silicon photonic continuous nonlinear unit and its experimentally measured transfer function. (a) False-color confocal micrograph of a fabricated neuron comprising a balanced photodetector pair and electro-optic modulator. (b-e) A variety of relevant O-E-O transfer functions seen from the PD-modulator pair, taken at different bias conditions: (b,c) sigmoids; (d,e) rectified linear units ("ReLU"). (f,g) Time resolved pictures of transfer functions: (f) the input is a 40-ns burst of a 100-MHz carrier; (g) both ReLUs. Figure and caption adapted from [36], with permission.

characterized by a Lorentzian, proper restriction of the current swings allows emulation of popular deep learning activation functions such as rectified-linear (ReLu), sigmoid, and quadratic (not shown). Similar to the weighting case, electro-absorption modulators were also proposed for this purpose [37].

An important property of the nonlinear unit is its cascadability. Physical cascadability requires that the output of a neuron be compatible with its input i.e., that the nodes described be networked. The incoherent units of Figure 16.2 do this by matching the resonator modulator's wavelengths to the resonant filters used for the MAC operations. Gain cascadability quantifies if the output of a neuron suffices to drive all of its fan-out. The specific amount of gain required within a single nonlinear unit ultimately depends on signal levels and network characteristics such as fan-in, fan-out, and level of attenuation by the weights. Finally, for cascadability to be present from the point of view of noise, we require a signal-to-noise ratio >1 after the signal has propagated through the network. This means the neuronal nonlinearity must counteract the amplitude and phase noise due to e.g., imperfect control elements that are present in both coherent and incoherent approaches.

A quantified measure of cascadability can be obtained through an autapse, or self-connection, experiment. Such experiments have been performed in both laser spiking [38] and silicon modulator neurons [36]. For the forward-biased pn-junction modulator neuron of Figure 16.2, for instance, a minimum optical pump power of $2V_{\pi}/\pi R_{pd}R_b$ is required to have a gain larger than unity, and this is seen in experiments. The balance of signal degradation and noise trimming from the Lorentzian transmission in a pn-junction system with realistic component values was studied theoretically [39]. The results yielded over 50 dB of calculated signal-to-noise after an arbitrarily large network, hinting that such a system offers an amount of cascadability from the point of view of noise.

16.3 SILICON PHOTONIC NEURAL NETWORKS AND APPLICATIONS

Given a cascadable nonlinear unit (or layer) that can take in and perform weighted summation of input signals, neural networks proper can be considered. Coherent layers, by preserving their path-encoding scheme, are straightforwardly cascaded to form a deep feedforward network, although the mesh routing must be maintained [22]. For incoherent neurons where every node emits on its own channel (wavelength, mode, polarization), all of which can exist in the same single physical waveguide, the "broadcast-and-weight" protocol was introduced in [23]. The O/E/O conversion can be leveraged to isolate different waveguides on the same chip, enabling spectrum reuse. Which such "broadcast loop(s)" a neuron outputs into determines the network topology, with outputs back to a neuron's inputs allowing recurrence. The specific network instantiation given this physical topology, which neuron connects to which and how strongly, is then dictated by the synaptic weight described in Section 16.2.1. For microring-based neurons as previous-discussed, the microring resonator filters can be assembled into what are called weight banks as displayed in Figure 16.3a [40]. Temporal multiplexing can be used to go beyond limited amount of hardware. Fast implementation of convolutional neural networks has been proposed this way by [41,42].

An approach related to recurrent neural networks operating at the network level, called reservoir computing, is briefly mentioned here since it was also demonstrated in silicon photonics [43]. The idea is to create a network with (semi) random connections such that it exhibits nontrivial dynamics, send time-series data through the system, and train e.g., the output layer of this "reservoir" [44,45]. This is attractive since it is easier than training a full recurrent neural network. It is a popular approach for photonics in general [46–48]. While sharing with deep learning the training of network parameters conditional on data, the requirement on the "neurons" as described in the preceding section is relaxed in this case, since the only requirement of the network is that it "lifts" the time-series input to a higher-dimensional space, effectively performing feature extraction. For instance, in silicon photonics, interference in passive structures and nonlinearity from photodetection are found to be enough to successfully perform computing tasks [43,49].

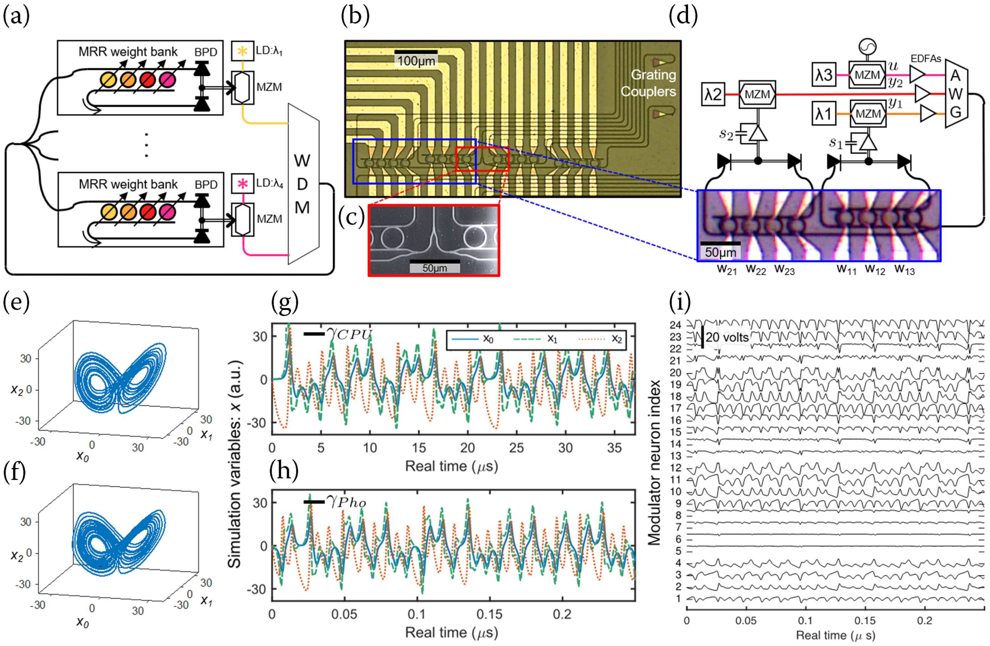

FIGURE 16.3 Photonic neural network benchmarking against a CPU. (a) Concept of a broadcast-and-weight network with modulators used as neurons. MRR: microring resonator, BPD: balanced photodiode, LD: laser diode, MZM: Mach-Zehnder modulator, WDM: wavelength-division multiplexer. (b) Micrograph of 4-node recurrent broadcast-and-weight network with 16 tunable microring (MRR) weights and fiber-to-chip grating couplers. (c) Scanning electron micrograph of 1:4 splitter. (d) Experimental setup with two off-chip MZM neurons and one external input. Signals are wavelength-multiplexed in an arrayed waveguide grating (AWG) and coupled into a 2 × 3 subnetwork with MRR weights, w_{11}, w_{12}, etc. Neuron state is represented by voltages s_1 and s_2 across low-pass filtered transimpedance amplifiers, which receive inputs from the balanced photodetectors of each MRR weight bank. (e,f) Phase diagrams of the Lorenz attractor simulated by a conventional CPU (e) and a photonic CTRNN (b). (g,h) Time traces of simulation variables for a conventional CPU (g) and a photonic CTRNN (h). The horizontal axes are labeled in physical real time, and cover equal intervals of virtual simulation time, as benchmarked by γ_{CPU} and γ_{Pho}. The ratio of real-time values of γs indicates a 294-fold acceleration. (i) Time traces of modulator voltages s_i (minor y-axis) for each modulator neuron i (major y-axis) in the photonic CTRNN. The simulation variables, x, in (h) are linear decodings of physical variables, s, in (i). Figure and caption adapted from [24], with permission.

16.3.1 Application I: Neural ODE Solver

With an interconnection strategy, networks can be created and actual processing tasks considered. For instance, using the neural compiler Nengo [50], incoherent silicon photonic modulator neurons in a broadcast-and-weight configuration were configured for time series prediction [24]. The neural network emulates three coupled ODEs (Lorenz attractor), with parameters set to produce a chaotic output. The physical device and experimental setup, reproduced in Figure 16.3a-d, is used to extract experimental behavior for a single node. A 24-node network is then emulated, and its performance is favorably benchmarked against a CPU solving the same problem, with a predicted speedup of 294x.

16.3.2 Application II: Nonlinear Programming and Model-Predictive Control

If a neural network can model time dynamics quickly, model-predictive control is an interesting next step. Model-predictive control is a nonlinear scheme that, in opposition to linear control such

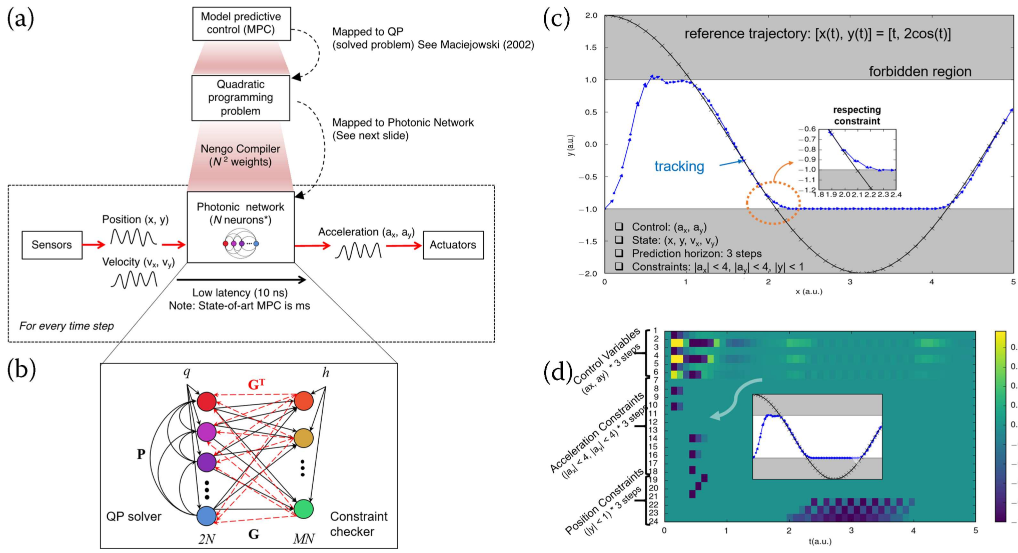
(a)
Model predictive control (MPC)
Mapped to QP (solved problem) See Maciejowski (2002)
Quadratic programming problem
Mapped to Photonic Network (See next slide)
Nengo Compiler (N^2 weights)
Sensors
Position (x, y)
Velocity (v_x, v_y)
Photonic network (N neurons*)
Acceleration (a_x, a_y)
Actuators
Low latency (10 ns)
Note: State-of-art MPC is ms
For every time step
(b)
q
h
G^T
P
QP solver
2N
G
MN
Constraint checker
(c)
reference trajectory: [x(t), y(t)] = [t, 2cos(t)]
forbidden region
tracking
respecting constraint
□ Control: (a_x, a_y)
□ State: (x, y, v_x, v_y)
□ Prediction horizon: 3 steps
□ Constraints: $|a_x| < 4$, $|a_y| < 4$, $|y| < 1$
y (a.u.)
x (a.u.)
(d)
Control Variables (ax, ay) * 3 steps
Acceleration Constraints ($|a_x| < 4$, $|a_y| < 4$) * 3 steps
Position Constraints ($|y| < 1$) * 3 steps
t(a.u.)
normalized neuron output

FIGURE 16.4 (a) Schematic figure of the procedure to implement the MPC algorithm on a neuromorphic photonic processor. Firstly, map the MPC problem to QP. Then, construct a QP solver with continuous-time recurrent neural networks (CT-RNN). Finally, build a neuromorphic photonic processor to implement the CT-RNN. (b) Schematic figure of construction of a QP solver with CT-RNN. In this example, $N = 3$, which is the prediction horizon, $M = 6$, which is the number of inequalities, and 2 is the vector dimension. (c) The trajectory of the moving target is shown in the black curve, and the blue dots and blue arrows are the simulated results of the position and velocity of the tracker at each time step respectively. The inset shows that the controller predicts a constraint violation and starts turning the tracker to avoid violating the acceleration's constraint. (d) The "constraint checker" neurons fire around $t = 0.5$ and between $t = 2$ and $t = 4$, inhibiting the output of the "QP solver neurons" such that the outcome of the system does not violate the acceleration and position constraints, respectively. Figure and caption adapted from [31], with permission.

as PID where only the immediate past is responded to, models future trajectories and proactively corrects its course towards an optimal solution that avoids constraints. Electronic processors implementing these algorithms for e.g., chemical plant process these are capped at kHz speeds, whereas photonic neural networks can potentially reach hundreds of MHz, enabling new control regimes [31]. Figure 16.4 shows this process. The task must first be mapped to a quadratic programming problem, which can then be compiled by a neural compiler into a neural network. The network contains a layer dedicated to solving the quadratic problem, and another to enforce the constraints. A photonic neural network can converge to a solution on the order of its neuron timescale, about 10 ns for the neuron of Figure 16.2. Figure 16.4c-d shows a set of 24 neurons solving the model-predictive problem, while predicting in advance if the constraints will be violated and reacting accordingly.

16.3.3 Application III: Intelligent Signal Processing

An interesting use case is when data is already in the optical (and analog) domain. Fiber nonlinearity compensation is such a situation. Figure 16.5 displays microring modulator silicon photonic neurons described in Section 16.2.2 used to learn the transmission characteristics of a 10,080 km trans-pacific transmission link. A two-layer feedforward network of such units is trained to compensate the nonlinear transmission impairment and achieves a Q-factor improvement of 0.51 dB. The results with the network trained with experimental data are only 0.06 dB from a numerical simulation with the same parameters. Since the neuron operates at comparable

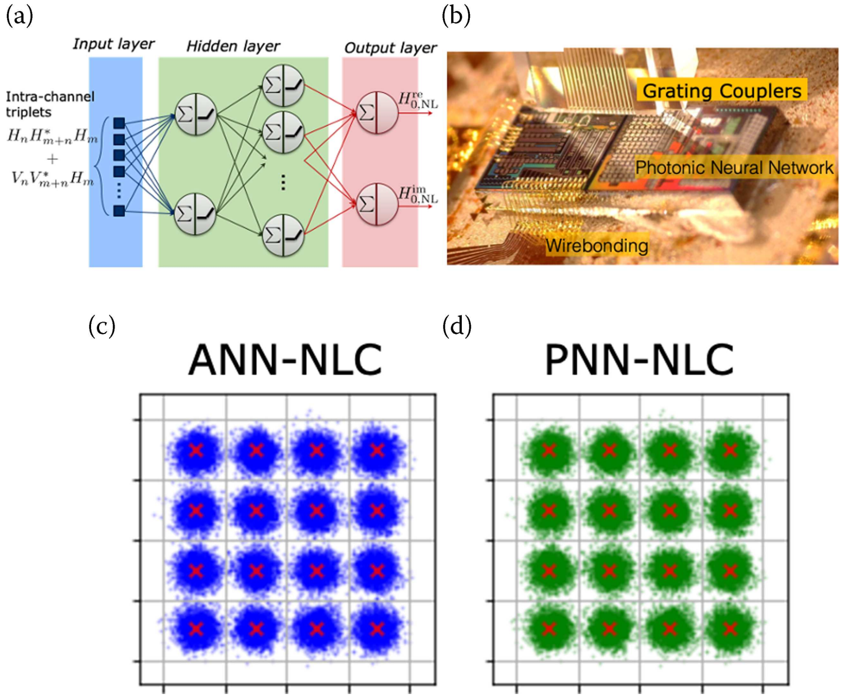

FIGURE 16.5 Photonic silicon photonic neural network for fiber nonlinearity compensation. (a) Schematic of ANN-NLC structure. (b) Image of the PNN chip under test and experimental setup for optical coupling and wirebonding. Constellations of X-polarization of a 32 Gbaud PM-16QAM, with the ANN-NLC gain of 0.57 dB in Q-factor (c) and with the PNN-NLC gain of 0.51 dB in Q-factor (d). Figure and caption adapted from [53], with permission.

bandwidths to the data rata, this suggests that real-time processing is within reach to improve optical communications.

Another intelligent signal processing application enabled by silicon neuromorphic photonic technology is in wideband RF signal processing. The linear part of silicon photonic neurons, for instance the weight banks displayed in Figure 16.3a-d, can be used on their own to perform e.g., principal component analysis [51] and independent component analysis [52]. These algorithms can perform dimensionality reduction on GHz modulated data in the optical domain to perform e.g., blind source separation. Doing this in an analog, wideband, passive photonic domain obviates the need for digital signal processing. It further requires only a single analog-digital conversion at the output instead of one per narrowband channel considered per input at the front end.

16.4 CONCLUSION AND FUTURE DIRECTIONS

Neuromorphic photonics has experienced rapid growth over the last few years. Still, there are many outstanding questions. The introduction of new electro-optic materials to silicon photonics would improve performance and functionality of silicon photonic neural networks. Tighter co-integration of light sources and amplifiers is a general aim of silicon photonics, and would also increase opportunities in neuromorphic engineering as well as ease of deployment [54]. The advent of zero-change CMOS silicon photonics platforms [55] is exciting to break the gain-bandwidth tradeoff in current silicon modulator neurons and improve density of control electronics. In the short term, this may be accomplished with separate dedicated CMOS chips wirebonded or flip-chip bonded to the silicon photonic chips. The applications reviewed were focused on inference tasks, but on-chip learning is also a very promising area. Spike timing–dependent plasticity was demonstrated with phase-change materials [26], and in the coherent approach, time-reversal symmetry to obtain gradients from intensity measurements was suggested to perform in-situ backpropagation [56]. In any case, ongoing investigations in neuromorphic photonics enabled by silicon photonics promise to bring machine intelligence to unexplored regimes, a salutary direction for a society increasingly dependent on neural network processors.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[2] J. Schmidhuber, 'Deep learning in neural networks: An overview', *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[3] F. Rosenblatt, 'The perceptron: A perceiving and recognizing automaton', Jan. 1957. Accessed: Oct. 16, 2020. [Online]. Available: https://blogs.umass.edu/brain-wars/files/2016/03/rosenblatt-1957.pdf.

[4] A. L. Hodgkin and A. F. Huxley, 'A quantitative description of membrane current and its application to conduction and excitation in nerve', *J. Physiol.*, vol. 117, no. 4, pp. 500–544, 1952, doi: 10.1113/jphysiol.1952.sp004764.

[5] C. Mead and M. Ismail, Eds., *Analog VLSI Implementation of Neural Systems*. Springer US, 1989.

[6] 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, doi: 10.1007/s11227-011-0631-3.

[7] N. P. Jouppi *et al.*, 'In-datacenter performance analysis of a tensor processing unit', *ArXiv170404760 Cs*, Apr. 2017, Accessed: Oct. 24, 2020. [Online]. Available: http://arxiv.org/abs/1704.04760.

[8] J. Duarte *et al.*, 'Fast inference of deep neural networks in FPGAs for particle physics', *J. Instrum.*, vol. 13, no. 07, p. P07027, Jul. 2018, doi: 10.1088/1748-0221/13/07/P07027.

[9] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, 'Physics for neuromorphic computing', *Nat. Rev. Phys.*, vol. 2, no. 9, pp. 499–510, Sep. 2020, doi: 10.1038/s42254-020-0208-2.

[10] K. Berggren *et al.*, 'Roadmap on emerging hardware and technology for machine learning', *Nanotechnology*, vol. 32, no. 1, p. 012002, Jan. 2021, doi: 10.1088/1361-6528/aba70f.

[11] D. Psaltis and N. Farhat, 'Optical information processing based on an associative-memory model of neural nets with thresholding and feedback', *Opt. Lett.*, vol. 10, no. 2, pp. 98–100, Feb. 1985, doi: 10.1364/OL.10.000098.

[12] L. Chrostowski and M. Hochberg, *Silicon Photonics Design: From Devices to Systems*. Cambridge: Cambridge University Press, 2015.

[13] P. R. Prucnal and B. J. Shastri, *Neuromorphic Photonics*. CRC Press, 2017.

[14] L. D. Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli, 'Photonic neural networks: A survey', *IEEE Access*, vol. 7, pp. 175827–175841, 2019, doi: 10.1109/ACCESS.2019.2957245.

[15] T. F. de Lima *et al.*, 'Primer on silicon neuromorphic photonic processors: Architecture and compiler', *Nanophotonics*, vol. 9, no. 13, pp. 4055–4073, Aug. 2020, doi: 10.1515/nanoph-2020-0172.

[16] X. Sui, Q. Wu, J. Liu, Q. Chen, and G. Gu, 'A review of optical neural networks', *IEEE Access*, vol. 8, pp. 70773–70783, 2020, doi: 10.1109/ACCESS.2020.2987333.

[17] B. J. Shastri *et al.*, 'Photonics for artificial intelligence and neuromorphic computing', *ArXiv201100111 Phys.*, Oct. 2020, Accessed: Nov. 05, 2020. [Online]. Available: http://arxiv.org/abs/2011.00111.

[18] D. Rosenbluth, K. Kravtsov, M. P. Fok, and P. R. Prucnal, 'A high performance photonic pulse processing device', *Opt. Express*, vol. 17, no. 25, pp. 22767–22772, Dec. 2009, doi: 10.1364/OE.1 7.022767.

[19] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, 'All-optical spiking neurosynaptic networks with self-learning capabilities', *Nature*, vol. 569, no. 7755, pp. , 208–214, May 2019, doi: 10.1038/s41586-019-1157-8.

[20] J. M. Shainline, S. M. Buckley, R. P. Mirin, and S. W. Nam, 'Superconducting optoelectronic circuits for neuromorphic computing', *Phys. Rev. Appl.*, vol. 7, no. 3, p. 034013, Mar. 2017, doi: 10.1103/PhysRevApplied.7.034013.

[21] M. A. Nahmias, T. F. de Lima, A. N. Tait, H. Peng, B. J. Shastri, and P. R. Prucnal, 'Photonic multiply-accumulate operations for neural networks', *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1–18, Jan. 2020, doi: 10.1109/JSTQE.2019.2941485.

[22] Y. Shen *et al.*, 'Deep learning with coherent nanophotonic circuits', *Nat. Photonics*, vol. 11, no. 7, pp. 441–446, Jul. 2017, doi: 10.1038/nphoton.2017.93.

[23] 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. Light. Technol.*, vol. 32, no. 21, pp. 3427–3439, Nov. 2014.

[24] A. N. Tait *et al.*, 'Neuromorphic photonic networks using silicon photonic weight banks', *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, Aug. 2017, doi: 10.1038/s41598-017-07754-z.

[25] M. Miscuglio and V. J. Sorger, 'Photonic tensor cores for machine learning', *Appl. Phys. Rev.*, vol. 7, no. 3, p. 031404, Jul. 2020, doi: 10.1063/5.0001942.

[26] Z. Cheng, C. Ríos, W. H. P. Pernice, C. D. Wright, and H. Bhaskaran, 'On-chip photonic synapse', *Sci. Adv.*, vol. 3, no. 9, p. e1700160, Sep. 2017, doi: 10.1126/sciadv.1700160.

[27] C. Ríos *et al.*, 'In-memory computing on a photonic platform', *Sci. Adv.*, vol. 5, no. 2, p. eaau5759, Feb. 2019, doi: 10.1126/sciadv.aau5759.

[28] A. N. Tait *et al.*, 'Feedback control for microring weight banks', *Opt. Express*, vol. 26, no. 20, pp. 26422–26443, Oct. 2018, doi: 10.1364/OE.26.026422.

[29] C. Huang *et al.*, 'Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits', *APL Photonics*, vol. 5, no. 4, p. 040803, Apr. 2020, doi: 10.1063/1.5144121.

[30] F. Akopyan *et al.*, 'TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip', *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015, doi: 10.1109/TCAD.2015.2474396.

[31] T. F. de Lima *et al.*, 'Machine learning with neuromorphic photonics', *J. Light. Technol.*, vol. 37, no. 5, pp. 1515–1534, Mar. 2019, doi: 10.1109/JLT.2019.2903474.

[32] W. Zhang, R. Mazzarello, M. Wuttig, and E. Ma, 'Designing crystallization in phase-change materials for universal memory and neuro-inspired computing', *Nat. Rev. Mater.*, vol. 4, no. 3, pp. 150–168, Mar. 2019, doi: 10.1038/s41578-018-0076-x.

[33] C. Huang, A. Jha, T. F. de Lima, A. N. Tait, B. J. Shastri, and P. R. Prucnal, 'On-chip programmable nonlinear optical signal processor and its applications', *IEEE J. Sel. Top. Quantum Electron.*, vol. 27, no. 2, pp. 1–11, Mar. 2021, doi: 10.1109/JSTQE.2020.2998073.

[34] A. Jha, C. Huang, and P. R. Prucnal, 'Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics', *Opt. Lett.*, vol. 45, no. 17, pp. 4819–4822, Sep. 2020, doi: 10.1364/OL.398234.

[35] M. M. P. Fard *et al.*, 'Experimental realization of arbitrary activation functions for optical neural networks', *Opt. Express*, vol. 28, no. 8, pp. 12138–12148, Apr. 2020, doi: 10.1364/OE.391473.

[36] A. N. Tait *et al.*, 'Silicon photonic modulator neuron', *Phys. Rev. Appl.*, vol. 11, no. 6, p. 064043, Jun. 2019, doi: 10.1103/PhysRevApplied.11.064043.

[37] R. Amin *et al.*, 'ITO-based electro-absorption modulator for photonic neural activation function', *APL Mater.*, vol. 7, no. 8, p. 081112, Aug. 2019, doi: 10.1063/1.5109039.

[38] H.-T. Peng *et al.*, 'Autaptic circuits of integrated laser neurons', in *Conference on Lasers and Electro-Optics (2019), paper SM3N.3*, May 2019, p. SM3N.3, doi: 10.1364/CLEO_SI.2019.SM3N.3.

[39] T. F. de Lima *et al.*, 'Noise analysis of photonic modulator neurons', *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1–9, Jan. 2020, doi: 10.1109/JSTQE.2019.2931252.

[40] A. N. Tait *et al.*, 'Microring weight banks', *IEEE J. Sel. Top. Quantum Electron.*, vol. 22, no. 6, pp. 312–325, Nov. 2016, doi: 10.1109/JSTQE.2016.2573583.

[41] A. Mehrabian, Y. Al-Kabani, V. J. Sorger, and T. El-Ghazawi, 'PCNNA: A photonic convolutional neural network accelerator', in *2018 31st IEEE International System-on-Chip Conference (SOCC)*, Sep. 2018, pp. 169–173, doi: 10.1109/SOCC.2018.8618542.

[42] V. Bangari *et al.*, 'Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs)', *ArXiv190701525 Phys.*, Apr. 2019, Accessed: Oct. 30, 2020. [Online]. Available: http://arxiv.org/abs/1907.01525.

[43] K. Vandoorne *et al.*, 'Experimental demonstration of reservoir computing on a silicon photonics chip', *Nat. Commun.*, vol. 5, no. 1, pp. 1–6, Mar. 2014, doi: 10.1038/ncomms4541.

[44] W. Maass, T. Natschläger, and H. Markram, 'Real-time computing without stable states: A new framework for neural computation based on perturbations', *Neural Comput.*, vol. 14, no. 11, pp. 2531–2560, Nov. 2002, doi: 10.1162/089976602760407955.

[45] H. Jaeger and H. Haas, 'Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication', *Science*, vol. 304, no. 5667, pp. 78–80, Apr. 2004, doi: 10.1126/science.1091277.

[46] 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, no. 1, pp. 1–7, Jan. 2013, doi: 10.1038/ncomms2368.

[47] L. Appeltant *et al.*, 'Information processing using a single dynamical node as complex system', *Nat. Commun.*, vol. 2, no. 1, pp. 1–6, Sep. 2011, doi: 10.1038/ncomms1476.

[48] L. Larger *et al.*, 'Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing', *Opt. Express*, vol. 20, no. 3, pp. 3241–3249, Jan. 2012, doi: 10.1364/OE.20.003241.

[49] A. Katumba *et al.*, 'Neuromorphic computing based on silicon photonics and reservoir computing', *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 6, pp. 1–10, Nov. 2018, doi: 10.1109/JSTQE.2018.2821843.

[50] T. Bekolay *et al.*, 'Nengo: A Python tool for building large-scale functional brain models', *Front. Neuroinformatics*, vol. 7, Jan. 2014, doi: 10.3389/fninf.2013.00048.

[51] A. N. Tait *et al.*, 'Demonstration of multivariate photonics: Blind dimensionality reduction with integrated photonics', *J. Light. Technol.*, vol. 37, no. 24, pp. 5996–6006, Dec. 2019, doi: 10.1109/JLT.2019.2945017.

[52] P. Y. Ma *et al.*, 'Photonic independent component analysis using an on-chip microring weight bank', *Opt. Express*, vol. 28, no. 2, pp. 1827–1844, Jan. 2020, doi: 10.1364/OE.383603.

[53] C. Huang *et al.*, 'Demonstration of photonic neural network for fiber nonlinearity compensation in long-haul transmission systems', in *Optical Fiber Communication Conference Postdeadline Papers 2020 (2020), paper Th4C.6*, Mar. 2020, p. Th4C.6, doi: 10.1364/OFC.2020.Th4C.6.

[54] D. Liang and J. E. Bowers, 'Recent progress in lasers on silicon', *Nat. Photonics*, vol. 4, no. 8, pp. 511–517, Aug. 2010, doi: 10.1038/nphoton.2010.167.

[55] V. Stojanović *et al.*, 'Monolithic silicon-photonic platforms in state-of-the-art CMOS SOI processes [Invited]', *Opt. Express*, vol. 26, no. 10, pp. 13106–13121, May 2018, doi: 10.1364/OE.26.013106.

[56] 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, pp. 864–871, Jul. 2018, doi: 10.1364/OPTICA.5.000864.