An integrated analog O/E/O link for multi-channel laser neurons

Mitchell A. Nahmias, Alexander N. Tait, Leonidas Tolias, Matthew P. Chang, Thomas Ferreira de Lima, Bhavin J. Shastri, and Paul R. Prucnal

Citation: Appl. Phys. Lett. **108**, 151106 (2016); doi: 10.1063/1.4945368
View online: http://dx.doi.org/10.1063/1.4945368
View Table of Contents: http://aip.scitation.org/toc/apl/108/15
Published by the American Institute of Physics

An integrated analog O/E/O link for multi-channel laser neurons

Mitchell A. Nahmias,[a] Alexander N. Tait, Leonidas Tolias, Matthew P. Chang, Thomas Ferreira de Lima, Bhavin J. Shastri, and Paul R. Prucnal
Electrical Engineering Department, Princeton University, 41 Olden St, Princeton, New Jersey 08540, USA

(Received 9 January 2016; accepted 23 March 2016; published online 14 April 2016)

We demonstrate an analog O/E/O electronic link to allow integrated laser neurons to accept many distinguishable, high bandwidth input signals simultaneously. This device utilizes wavelength division multiplexing to achieve multi-channel fan-in, a photodetector to sum signals together, and a laser cavity to perform a nonlinear operation. Its speed outpaces accelerated-time neuromorphic electronics, and it represents a viable direction towards scalable networking approaches. © *2016 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4945368]

Optical signals exhibit many advantages over their electronic counterparts—they consume less energy per bit, have a greater bandwidth density per wire, and display lower cross-talk between multiplexed channels. These physical properties have spawned a renewed interest in the computational abilities of optical systems. Non-traditional computing architectures such as reservoir computing have shown significant performance gains.[1–3] These improvements stem largely from the enormous bandwidth of optical waveguides, which, for a typical telecommunications band, is around several terahertz. Likewise, laser cavities have also been investigated for their ultrafast dynamical properties, which can be exploited for nonlinear processing. Many recent investigations have focused on neuromorphic models,[4–8] which can operate many millions ($>10^7$) of times faster than their biological counterparts.

Nonetheless, networking neurons can be challenging in hardware. Since the number of interconnects scales with N^2 as a function of the number of neurons N in a fully connected network, electronic approaches utilize time multiplexing strategies to allow for shared buses that prevent wire counts from becoming unmanageable. The most popular of these is address event representation, which assigns a unique packet signature to each neuron.[9] Unfortunately, these techniques require multiplexers and demultiplexers that operate at a much higher bandwidth than that of the network itself. As a result, the fastest spiking neural networks have temporal resolutions in the 10s of MHz,[10] and the bandwidth limitations of electronic wires may prevent networks from ever reaching gigahertz speeds.

The enormous bandwidth of optical waveguides may help alleviate the speed limitations of electronic approaches. The ability to perform weighted addition (i.e., $x_\Sigma = \sum_i w_i x_i$ for inputs x_i) would allow each node to receive multiple signals simultaneously. This operation—which can be viewed as a parallelized multiply-and-accumulate—is crucial for complex computations and also costly to implement in hardware neural network models.[11] Although researchers have shown a plethora of nonlinear effects in semiconductor lasers, none of these models have demonstrated the ability to accept multiple inputs. As a result, networks of only a few neurons have been studied theoretically,[4,8] and the ability to receive many inputs has not been demonstrated in a multi-neuron experiments.[12,13] In a response, a scalable networking scheme based on wavelength division multiplexing (WDM) was proposed,[14] which addressed many of the issues of cascadability and fan-in for networks of ultrafast spiking laser neurons. However, this scheme requires wavelength conversion between input and output, creating a challenge for laser injection approaches.

In this paper, we experimentally demonstrate a semiconductor laser neuron model that uses WDM to solve the networking bottleneck and perform weighted addition of input signals. In this architecture, a photodetector (PD) receives optical inputs and drives an adjacent laser, which performs a nonlinear operation and converts the signal back into the optical domain. The key design innovation is a direct analog O/E/O link that drives a laser, requiring no additional circuitry to process the signal.[15] This conversion is conventionally associated with a high performance cost to demultiplex, digitize, and remodulate signals. However, in an analog neural processing context, the electronic link can enable wavelength conversion without necessarily performing these costly functions. Devices of this sort could be compatible with all-to-all WDM networks (Figs. 1(a)–1(c)), free from the bandwidth limitations of electronic interconnects.[14] This configuration has been explored in a fiber prototype[12] and has architectural similarities with Ref. 6. The system would be compatible with more complex dynamical operations, including a variety of the models described in Ref. 16. The device has a bandwidth of several hundred MHz and has the potential to go up to ~10s of GHz,[17] outpacing electronic approaches by several orders of magnitude.

This experiment explores an O/E/O link with a continuous perceptron-like neuron model,[18] which is simpler than more recent spike-based models explored in the hardware community[11] and in computational neuroscience.[19] A simple neuron model includes two primary operations: a weighted sum of input signals (i.e., $x = \sum_i w_i s_i$ for weights w_i and inputs s_i) and a nonlinear operation (i.e., $y = f\{x\}$ for some nonlinear filter f). The nonlinearity can vary from a simple sigmoid to a complex dynamical system, depending on the neural model used.[20] Spiking models, which are based on more complex nonlinear dynamics, promise better energy

[a]Electronic mail: mnahmias@princeton.edu

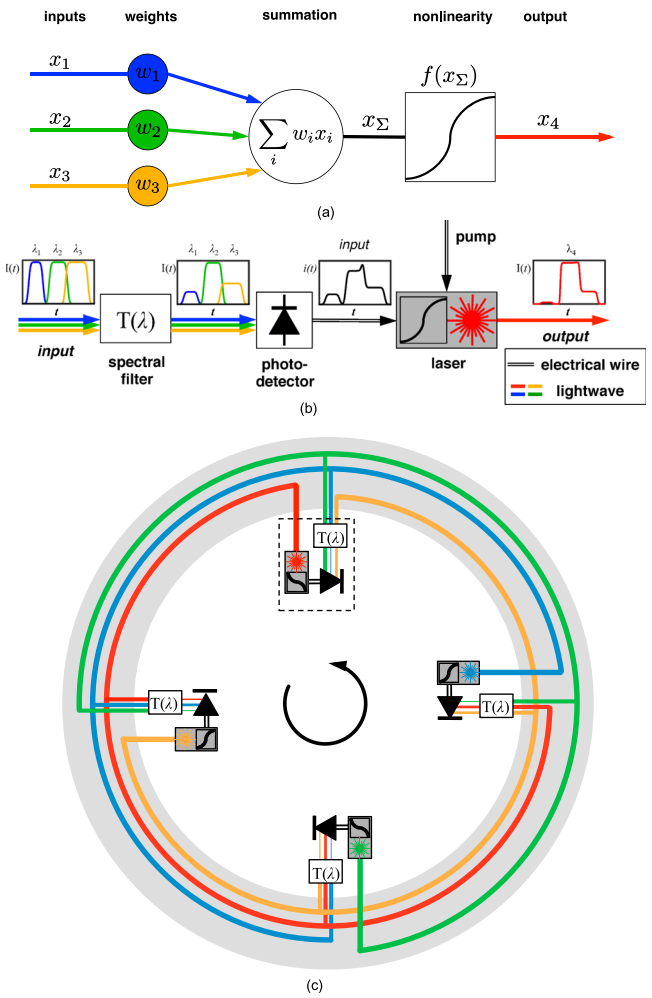

FIG. 1. (a) Schematic of a neural network node. Inputs x_1, x_2, x_3 are weighted and summed s.t. $x_\Sigma = \sum_i w_i x_i$, and then experience a nonlinear function $y = f(x_\Sigma)$. (a) Concept diagram for a processing network node (PNN). Inputs incident from other lasers at different wavelengths $\lambda_{1,2,3}$ are spectrally filtered (i.e., weighted). This is followed by conversion of the total power into an electronic signal (i.e., summation) by a photodetector (PD), which drives a laser performing a nonlinear operation. The laser is pumped with an electronic direct current signal. The output at a new wavelength λ_4 feeds back into the network. (b) Diagram of a simple network topology made possible by this architecture, where the black arrow indicates the direction of light propagation. The PNN above is boxed in a dotted line. Using this architecture, neurons can form all-to-all networks by sharing a single broadcast waveguide.[14]

efficiency through sparse coding and the ability to operate close to the noise floor.[21] It is also the focus of recent research in the exploration of photonic processors.[16] However, spiking behavior is considerably more complex,[20] and as a result, more difficult to emulate in hardware. In addition, there are fewer known algorithms that can utilize spike codes efficiently compared to more traditional approaches.[22] Simple continuous-time perceptron models, on the other hand, can approximate any function,[23] simulate any dynamical system,[24] and form the basis of high-performance machine algorithms.[25] Nonetheless, the O/E/O driving framework is amenable to spiking laser models, provided they can be modulated with a current signal.[16]

Active devices were fabricated using a standard AlGaInAs multi-quantum well epitaxial structure on indium phosphide, designed to operate in the optical C-band. 5 μm width waveguide ridges were fabricated using a wet etch technique. This was followed by an electrical insulation layer (SiO_2) through plasma enhanced chemical vapor deposition, dry etching of contact windows, and metal deposition. Samples were lapped to a 50 μm thickness before being manually cleaved and mounted onto custom chip-scale submounts. Submounts were fabricated on 1 cm^2 silicon chips: an oxide layer (SiO_2) with a thickness of ~50 nm was deposited on each sample to provide insulation, followed by metal deposition. Our samples were positioned by hand and mounted to the grounding pads via silver epoxy. This was followed by aluminum wire-bonding between the PD and laser. Fig. 2 shows a picture of the fully fabricated device.

Two fiber tapers were aligned to the PD and the laser independently at orthogonal angles (Fig. 2). The input was generated using a configuration similar to Ref. 26, which consisted of a pulse pattern generator (PPG) and an array of variable optical attenuators (VOAs) and delay lines nested between two arrayed waveguide gratings (AWG). Since only a single PPG signal was used for all wavelength channels,

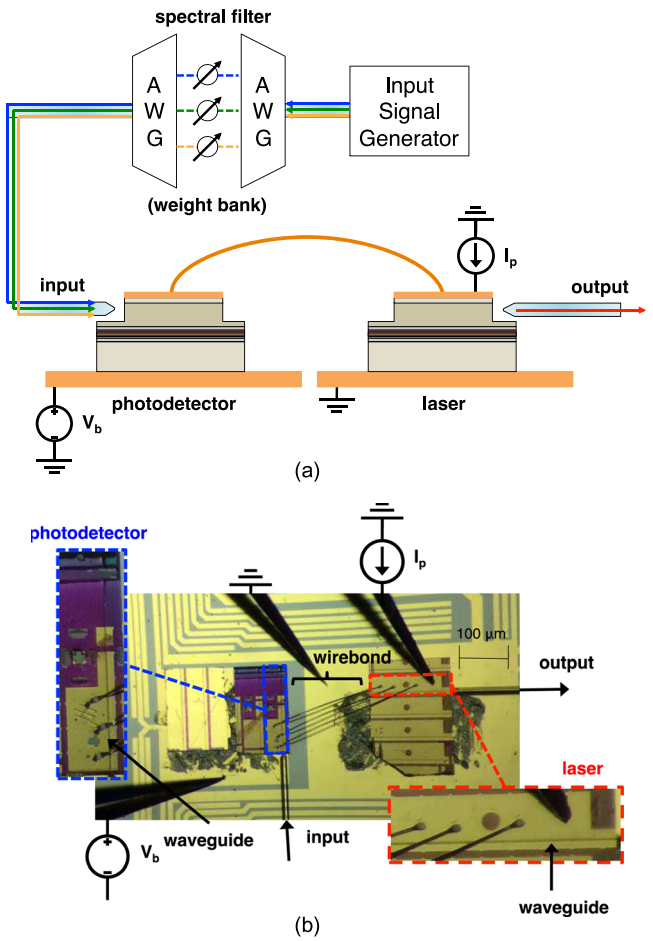

FIG. 2. (a) Experimental set-up, mirroring the basic processing node schematic (Fig. 1(b)). Generated input signals experience different weights through a series of variable attenuators nested between two arrayed waveguide gratings (AWGs). The WDM signal travels into the PD, which current modulates an adjacent laser. The laser output is subsequently measured on a sampling scope. (b) Picture of a fabricated device, taken under a microscope. Active InGaAlAs device is mounted on separate grounding pads, fabricated on a silicon submount. The top (P-side) of the PD is connected directly to the top (P-side) of the laser with aluminum wire bonds. With proper biasing ($V_b = 3$ V, I_p = 85–95 mA), optical inputs are converted into current signals, which a directly modulated laser converts back to an optical signal representing a thresholded version of the weighted sum.

delay lines were necessary to create independent signals on each channel. The generated signal was power monitored by an external high-frequency PD and oscilloscope system for input calibration. The input signal was amplified through an erbium doped fiber amplifier before arriving at the receiving PD. The resulting output was then measured using a PD connected to a high frequency sampling scope. Individual wavelength channels were measured by blanking out other channels via the VOA. A simple schematic of the experimental set-up (amplifier and oscilloscope not included) is shown in Fig. 2(a).

Since the laser and PD shared the same top contact, proper biasing required that separate ground pads rest under each component. A positive bias ($V_b = 3$ V) on the photodetector pad ensured reverse bias operation in the PD and counteraction of the forward pump current of the laser (85–95 mA) traveling across the wirebond junction. This biasing configuration, in addition to allowing the PD and laser to share the same top contact, ensured that the photocurrent signal traveled directly into the laser. As shown in Figs. 3(i)–3(ii), our composite device can receive signals from different wavelength channels simultaneously. The output of the laser is a simple bandwidth-limited sum of inputs. The weighting part of the operation $w_i x_i$, though not shown here, has been demonstrated using the same nested AWG system,[26] and has also been shown in an integrated silicon microring platform.[27]

Both operations of a simple neuron—a weighted sum of input signals ($x = \sum_i w_i s_i$) and a nonlinear operation ($y = f\{x\}$)—are demonstrated in this device. This experimental demonstration bridges both summation and nonlinear operations together using a simple PD-driving principle. Here, the nonlinear model is a simple threshold function (i.e., a perceptron), made possible through the laser's nonlinear relationship between light and intensity, i.e., it is L-I curve. The model only includes a simple L-I nonlinearity, but the PD-laser driving concept could apply to other lasers with richer dynamics, as explored in Ref. 16.

The processing node can operate in both a linear and nonlinear mode, depending on the applied bias as shown in Fig. 3(iv). This heterogeneity is useful, for example, in constructing well-defined optimization functions with constraints.[28] In the linear region of the L-I curve, the laser is pumped above threshold ($I_p = 95$ mA), and the output is a linear function of the input. It can be seen that the short, low amplitude pulses are further suppressed by the bandwidth limitations of the device. In the nonlinear region ($I_p = 85$ mA), lower amplitude (i.e., shorter pulse) signals fail to breach the lasing threshold and do not produce a corresponding output. This thresholding property is important for the cascadability of feedforward networks, which requires that amplitude noise decreases during every iteration.[29] In this case of a simple sigmoid, this condition can be met if signals below a given threshold are significantly suppressed and thus do not propagate forward. Pulse coincidence between both channels (occurring around 43 ns) results in a larger amplitude, a demonstration of signal summation of multiple wavelength channels. Linear summation occurs as a result of the device being biased within the linear region of the L-I curve (Fig. 3, right).

The bandwidth of the device (~250 MHz) is limited primarily by the photodetector in this demonstration. Nevertheless, it still outpaces electronic approaches: only application specific circuits have been able to reach comparable levels,[30] while neuron-inspired approaches in field programmable gate arrays have, at best, been demonstrated at ~50–100 MHz signal bandwidths.[31] Photonic approaches may not be as miniaturizable as electronic implementations, since photonic devices are limited by the diffraction limit of light (~1 μm). Nevertheless, they are capable of much higher speeds and lower processing latencies. Although the metal bridge has an associated capacitance, simulations based on an equivalent circuit model suggest that the bandwidth is primarily limited by device parasitics, which can be improved by shrinking the contact pads or size of devices. The bandwidth can be >10 GHz with optimized device geometries, as shown in Ref. 17. In addition, although the current implementation involves a direct wire-bond between the PD and laser, top side ground pads would allow the PD-laser system to be printed monolithically, either in a III-V or hybrid silicon/III-V platform.[15]

Several limitations must be investigated before such a device is scalable to larger systems. The first requirement to consider is cascadability: this includes the maintenance of signal integrity (i.e., signal-to-noise ratio) and the power required to cascade a signal from the output to input. Although the requirements for cascadability in spiking

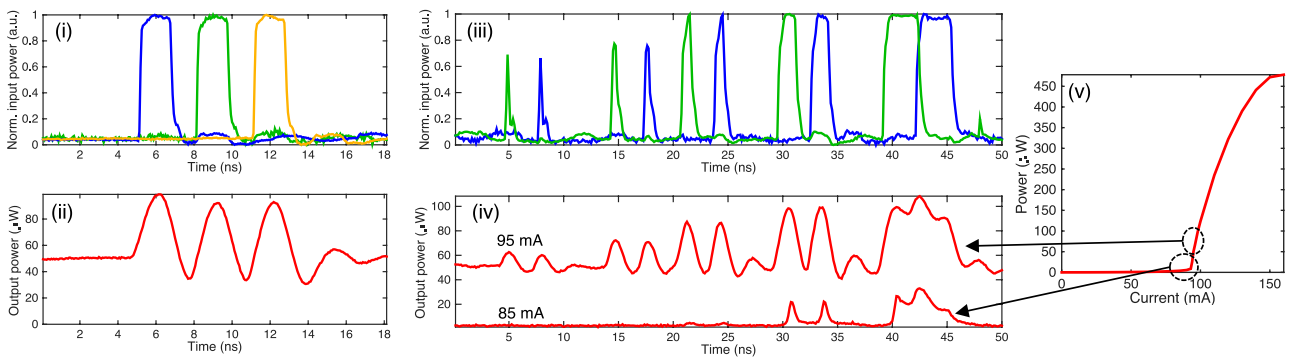

FIG. 3. Left: demonstration of multi-channel summation. (i) Normalized temporal pulse profiles of three independent wavelength channels (1538 nm, 1543 nm, 1548 nm) traveling into the photodetector. Input peak optical power is ~2 mW. (ii) The resulting output of the laser when biased at 95 mA. Right: demonstration of summation and thresholding. (iii) Inputs to the device, which include two wavelength channels (1538 nm, 1543 nm). Input peak optical power is ~7 mW. (iv) Output of laser biased below and above the lasing threshold. The laser can operate in both linear (95 mA) and nonlinear (85 mA) modes by using different regions of the L-I curve. If operated below threshold, the laser suppresses low-level signals, performing a key nonlinear operation necessary for processor cascadability. (v) The corresponding areas of the L-I curve used by the laser during operation.

models can be quite complex,[12] threshold-like models must simply suppress analog noise—a criterion met by the nonlinear L-I curve. As described in Fig. 3, the output optical power is reduced by nearly 20 dB. This can be attributed to free-space, polarization-dependent coupling loss, mirror loss, and the low energy efficiency of the laser. These aspects can be significantly improved with more optimized device geometries with an on-chip waveguide network. Nonetheless, several amplifications options exist to address this incongruity, including an electronic radio frequency (RF) amplifier between the PD and laser, and semiconductor optical amplifiers within the network. Electronic RF amplifiers can exhibit high gain (~20 dB) at high bandwidths (~60 GHz).[32] However, the use of these devices would require a platform that is compatible with both RF electronics and photonic components. Another option is to replace the laser with a modulator, which may reduce energy costs and increase input sensitivity. Regardless, the power budget must be carefully analyzed for full system integration.

Second, there are several scalability challenges that must be addressed for wavelength networking: filtering (i.e., the weighting mechanism) must be integrated on-chip, laser outputs must have narrow spectral linewidths, and the photodetectors must be able to receive many signals simultaneously. Although only three wavelengths were demonstrated here for clarity, the device shown here is capable of receiving many more—8× simultaneous input channels have been tested in experiment,[26] whereas the upper bound for theoretical channel count in silicon waveguide systems is approximately ~108.[33] To this end, there has been research in this area of using silicon microrings to perform weighted addition of wavelength-division multiplexed signals.[27] With regards to light generation, photonic crystal structures such as distributed feedback gratings are able to provide the narrow linewidths necessary.[15] Active and passive devices in III-V materials and silicon, respectively, can interface via adjacent butt coupling or fabricated lithographically in a hybrid III-V/silicon platform.[34]

In conclusion, we have fabricated a composite device structure that can be used as a processing node in a high bandwidth photonic neural network. The enormous bandwidth of optics is primarily exploited for network compatibility, allowing for a scalable fan-in approach. Although three wavelengths are demonstrated here for simplicity, scaling to larger numbers involves only the addition of more wavelengths and tunable filters. Networks that exceed wavelength channel limitations are also possible through the creation of multi-waveguide topologies, as investigated in Ref. 14. Our device demonstrates several important properties necessary for network integration and processing, including nonlinear thresholding and the summation of many distinguishable input signals (i.e., weighted addition) simultaneously. The device could serve as a template for future network-compatible laser neuron models, which could potentially operate at higher signal bandwidths with more rich, robust, or efficient laser dynamics.

M.A.N. and A.N.T. acknowledge the support of the National Science Foundation Graduate Research Fellowship. B.J.S. acknowledges the support of the Banting Postdoctoral Fellowship administered by the Government of Canada through the Natural Sciences and Engineering Research Council of Canada.

[1]K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, "Experimental demonstration of reservoir computing on a silicon photonics chip," Nat. Commun. **5**, 3541 (2014).
[2]A. Marandi, Z. Wang, K. Takata, R. L. Byer, and Y. Yamamoto, "Network of time-multiplexed optical parametric oscillators as a coherent Ising machine," Nat. Photonics **8**, 937–942 (2014).
[3]L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, "Information processing using a single dynamical node as complex system," Nat. Commun. **2**, 468 (2011).
[4]W. Coomans, L. Gelens, S. Beri, J. Danckaert, and G. Van der Sande, "Solitary and coupled semiconductor ring lasers as optical spiking neurons," Phys. Rev. E **84**, 036209 (2011).
[5]A. Hurtado, K. Schires, I. D. Henning, and M. J. Adams, "Investigation of vertical cavity surface emitting laser dynamics for neuromorphic photonic systems," Appl. Phys. Lett. **100**, 103703 (2012).
[6]B. Romeira, J. Javaloyes, C. N. Ironside, J. M. L. Figueiredo, S. Balle, and O. Piro, "Excitability and optical pulse generation in semiconductor lasers driven by resonant tunneling diode photo-detectors," Opt. Express **21**, 20931–20940 (2013).
[7]F. Selmi, R. Braive, G. Beaudoin, I. Sagnes, R. Kuszelewicz, and S. Barbay, "Relative refractory period in an excitable semiconductor laser," Phys. Rev. Lett. **112**, 183902 (2014).
[8]M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," IEEE J. Sel. Top. Quantum Electron. **19**, 1–12 (2013).
[9]K. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," IEEE Trans. Circuits Syst. II **47**, 416–434 (2000).
[10]J. Schemmel, D. Bruderle, A. Grubl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2010), pp. 1947–1950.
[11]J. Hasler and H. B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," Front. Neurosci. **7**, 118 (2013).
[12]B. J. Shastri, M. A. Nahmias, A. N. Tait, A. W. Rodriguez, B. Wu, and P. R. Prucnal, "Spike processing with a graphene excitable laser," Sci. Rep. **6**, 19126 (2016).
[13]T. V. Vaerenbergh, M. Fiers, P. Mechet, T. Spuesens, R. Kumar, G. Morthier, B. Schrauwen, J. Dambre, and P. Bienstman, "Cascadable excitability in microrings," Opt. Express **20**, 20292–20308 (2012).
[14]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. Lightwave Technol. **32**, 4029–4041 (2014).
[15]M. A. Nahmias, A. N. Tait, B. J. Shastri, T. Ferreira de Lima, and P. R. Prucnal, "Excitable laser processing network node in hybrid silicon: analysis and simulation," Opt. Express **23**, 26800–26813 (2015).
[16]P. R. Prucnal, B. J. Shastri, T. F. de Lima, M. A. Nahmias, and A. N. Tait, "Recent progress in semiconductor excitable lasers for photonic spike processing," Adv. Opt. Photonics (in press).
[17]M. A. Nahmias, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "A receiverless link for excitable laser neurons: Design and simulation," in *Summer Topicals Meeting Series (SUM), 2015* (IEEE, 2015) pp. 99–100.
[18]W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bull. Math. Biophys. **5**, 115–133 (1943).
[19]E. Izhikevich, "Simple model of spiking neurons," IEEE Trans. Neural Networks **14**, 1569–1572 (2003).
[20]W. Maass, "Networks of spiking neurons: The third generation of neural network models," Neural Networks **10**, 1659–1671 (1997).
[21]S. B. Laughlin, "Energy as a constraint on the coding and processing of sensory information," Curr. Opin. Neurobiol. **11**, 475–480 (2001).
[22]J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks **61**, 85–117 (2015).
[23]K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks **2**, 359–366 (1989).
[24]K.-i. Funahashi and Y. Nakamura, "Approximation of dynamical systems by continuous time recurrent neural networks," Neural Networks **6**, 801–806 (1993).
[25]Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).

[26]A. N. Tait, J. Chang, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, "Demonstration of WDM weighted addition for principal component analysis," Opt. Express **23**, 12758–12765 (2015).
[27]A. N. Tait, T. Ferreira De Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Continuous calibration of microring weights for analog optical networks," IEEE Photonic Technol. Lett. **28**(8), 887–890 (2016).
[28]M. Forti and A. Tesi, "New conditions for global stability of neural networks with application to linear and quadratic programming problems," IEEE Trans. Circuits Syst. I **42**, 354–366 (1995).
[29]D. A. B. Miller, "Are optical transistors the logical next step?" Nat. Photonics **4**, 3–5 (2010).
[30]P. Knag, J. K. Kim, T. Chen, and Z. Zhang, "A sparse coding neural network ASIC with on-chip learning for feature extraction and encoding," IEEE J. Solid-State Circuits **50**, 1070–1079 (2015).
[31]L. W. Kim, S. Asaad, and R. Linsker, "A fully pipelined FPGA architecture of a factored restricted Boltzmann machine artificial neural network," ACM Trans. Reconfigurable Technol. Syst. **7**, 5 (2014).
[32]J. Chen and A. M. Niknejad, "A compact 1V 18.6 dBm 60 GHz power amplifier in 65 nm CMOS," in *2011 IEEE International* Solid-*State Circuits Conference Digest of Technical Papers (ISSCC)* (2011), pp. 432–433.
[33]A. N. Tait, A. X. Wu, E. Zhou, T. Ferreira de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Microring weight banks: Simulation and performance analysis," J. Sel. Top. Quantum Electron. (unpublished).
[34]A. W. Fang, H. Park, Y.-H. Kuo, R. Jones, O. Cohen, D. Liang, O. Raday, M. N. Sysak, M. J. Paniccia, and J. E. Bowers, "Hybrid silicon evanescent devices," Mater. Today **10**, 28–35 (2007).