Broadcast and Weight: An Integrated Network For Scalable Photonic Spike Processing

Alexander N. Tait, *Student Member, OSA*, Mitchell A. Nahmias, Bhavin J. Shastri, *Member, IEEE*, and Paul R. Prucnal, *Fellow, IEEE, Fellow, OSA*

***Abstract*—We propose an on-chip optical architecture to support massive parallel communication among high-performance spiking laser neurons. Designs for a network protocol, computational element, and waveguide medium are described, and novel methods are considered in relation to prior research in optical on-chip networking, neural networking, and computing. Broadcast-and-weight is a new approach for combining neuromorphic processing and optoelectronic physics, a pairing that is found to yield a variety of advantageous features. We discuss properties and design considerations for architectures for scalable wavelength reuse and biologically relevant organizational capabilities, in addition to aspects of practical feasibility. Given recent developments commercial photonic systems integration and neuromorphic computing, we suggest that a novel approach to photonic spike processing represents a promising opportunity in unconventional computing.**

***Index Terms*—Asynchronous circuits, network topology, neuromorphics, optical computing, optical interconnects, photonic integrated circuits, spiking neural networks, system analysis and design, WDM networks.**

I. Introduction

NEUROMORPHIC processing offers many opportunities and challenges distinct from those of traditional von Neumann computing. It seeks to engineer scalable and cost-effective hardware systems that take inspiration from abstract principles of biological processing, such as parallelism and sparsity. Neuromorphic architectures promise potent advantages (efficiency, fault tolerance, adaptability) over von Neumann architectures for tasks involving pattern analysis, decision making, optimization, learning, and real-time control of multi-sensor, multi-actuator systems. Unconventional hardware has a long history of massive parallelism, but a more recently recognized point of neural inspiration is a sparse coding scheme called spiking [1].

Spike processing, while inspired by neuroscience, has firm code-theoretic justifications. Spike codes—digital in amplitude, but analog and sparse in pulse arrival time—can reconcile the expressiveness and efficiency of analog processing with the robustness of digital communication, and recurrent networks of spiking primitives possess rich algorithmic capabilities [2], [3]. Each spiking primitive handles inputs from multiple sources by temporally integrating their weighted sum and firing a single spike when the integration state variable crosses a threshold. This distributed, asynchronous model processes information using both space and time [4]–[6], and is amenable to distributed, unsupervised adaptation [7], [8]. The use of sparse coding principles promises extreme improvements to computational power efficiency in particular [9].

Spike processing is at the heart of a modern generation of neuromorphic electronics, although no single hardware approach has emerged as the clear ideal. Spiking primitives have been built in both CMOS analog circuits [10], digital "neurosynaptic cores" [11], and non-CMOS devices [12]. Many architectures that interconnect large numbers of primitives have been proposed or built, including, notably: Neurogrid [13], TrueNorth [14], SpiNNaker [15], and FACETS [16]. The use of physics for analog dynamical processing represents an important step towards attaining the efficiency and functionality exhibited by biophysical information processors, yet electronic interconnects are incapable of the density and fan-in needed to support scalable architectures that represent spikes as physical pulses. Despite a wide variety of approaches in neuromorphic microelectronics, all proposed architectures employ some form of address-event representation (AER) of spikes. AER is a digital packet routing scheme, which incurs significant time and energy overhead for signal encoding/decoding and network coordination, but is well-suited for slow timescale (milliseconds) neuromorphic systems [15].

Integrated photonic platforms represent an alternative to microelectronic approaches. The communication potentials of optical interconnects (bandwidth, energy use, electrical isolation) have received attention for neural networking in the past; however, attempts to realize holographic or matrix-vector multiplication systems have encountered practical barriers, largely because they cannot be integrated, let alone with effective nonlinear processing units. Techniques in silicon photonic integrated circuit (PIC) fabrication is driven by a tremendous demand for optical interconnects within conventional digital computing systems [17]. The first platforms for systems integration of active photonics are becoming commercial reality [18], [19], and promise to bring the economies of integrated circuit manufacturing to optical systems. Using a device set designed for digital communication (waveguides, filters, detectors, etc.), some have realized PICs for analog signal processing [20]. The potential of modern PIC platforms to enable large-scale all-optical systems for unconventional and/or analog computing has not yet been investigated.

Manuscript received February 14, 2014; revised May 27, 2014; accepted July 22, 2014. Date of publication August 5, 2014; date of current version September 17, 2014. This work was supported by the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP).

The authors are with the Lightwave Communications Laboratory, Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: atait@princeton.edu; mnahmias@princeton.edu; shastri@ieee.org; prucnal@princeton.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JLT.2014.2345652

Recent years have seen the emergence of a new class of optical devices that exploit a dynamical isomorphism between semiconductor photocarriers and neuron biophysics. The difference in physical timescales allows these "photonic neurons" to exhibit spiking behavior on picosecond (instead of millisecond) timescales [21]–[24]. Spiking is closely related to a dynamical system property called excitability, which is shared by certain kinds of laser devices. Excitable laser systems have been studied in the context of spike processing with the tools of bifurcation theory by [25]–[27] and experimentally by [28]–[30]. Some are specifically designed for compatibility with silicon photonic PIC platforms [31], [32]. A network of photonic neurons could open computational domains that demand unprecedented temporal precision, power efficiency, and functional complexity, potentially including applications in wideband radio frequency (RF) processing, adaptive control of multi-antenna systems, and high-performance scientific computing. Although the ultrafast spiking dynamics of laser neurons show potential in this respect, most analysis of them has so far been limited to one or two devices with minimal regard for a compatible network architecture.

We propose an on-chip networking architecture called "broadcast-and-weight" that could support massively parallel interconnection between photonic spiking neurons [33]. It has similarities with the fiber networking technique broadcast-and-select, which channelizes usable bandwidth using wavelength division multiplexing (WDM); however, the protocol flattens the traditional layered hierarchy of optical networks, accomplishing physical, logical, and processing tasks in a compact computational primitive. Although the proposed processing circuits are unconventional, the required device set is compatible with mainstream PIC platforms in silicon, which make heavy use of WDM techniques.

This paper is organized to first give background on optical networks on chip (NoC), computing, and neural networks. We will describe the WDM broadcast-and-weight protocol, then a primitive node for processing and networking, and a waveguide loop medium. Architectures consisting of multiple broadcast cells will be proposed and discussed with respect to topology and scalability. We have found that the implications of spike processing (time as information) combined with WDM (wavelength as identity) are accompanied by novel spatial freedoms that makes this architecture uniquely suited, among artificial systems, to emulate and explore certain biologically-relevant organizational topologies (e.g., "small-worldness"). This pairing also yields key features of practical feasibility (robustness, cascadability, scalability), which have foiled some large-scale optical processors in the past. We claim that various favorable and generalizable properties of the proposed architecture make it a viable candidate to support a new generation of scalable high-performance spike processing in photonics.

II. Prior Work

A. Optical Networks On-Chip

Optical networks on-chip (NoCs) have been proposed as an alternative to electronic networks to support the demanding throughput and efficiency requirements of future multi-core system on-chip architectures. Although the proposed interconnect is adapted for a considerably different signalling model (spiking), some of the networking techniques presented in this paper have been investigated in a conventional computing context. Optical ring networks with WDM channelization have been proposed as a means to obtain collision-free multicast networks, notably ATAC [34] and optical ring NoC (ORNoC) [35]. Psota *et al.* have also identified lightpath splitting as an efficient method for multicast routing on chip. The layout flexibility of the ring has been exploited to accommodate a tiled processor layout, and Le Beux *et al.* have proposed using multiple independent rings for spectrum reuse; however, interfacing these ORNoC subnetworks into a single system would require specialized switching nodes incorporating arbitration control, unlike the proposed architecture (see Section III-C).

WDM techniques significantly increase the effective throughput-density of a physical link; however, the requirement of a modulator and detector for each channel can negate the area and energy savings in some circumstances [36]. To obtain contention-free behavior, ATAC and ORNoC stipulate at least one dedicated receiver (i.e., detector, A/D converter, deserializer, and buffer) per channel per node, potentially creating a buffering bottleneck [35]. In contrast, the photonic spike processing architecture sums multiple inputs in a single detector and requires neither active electronic receivers nor distinct optical modulators (see Section III-B).

B. Optical Computing

Motivated by the properties that have made optics superior for communications (e.g., usable bandwidth and energy efficiency), optical devices and architectures have long been investigated for computing. Optical logic gates have been implemented by myriad techniques, including self-phase modulation in microring cavities [37], quantum dot saturable absorption [38], and many others; however, a scalable all-optical computer has so far proven elusive. Analyses of the daunting scope of fundamental challenges to digital optical computing are performed by Keyes [39] and Miller [40]. A comparison of these references reveals strikingly similar themes, which belie the progress of photonic technology in the intervening decades – not to mention the birth and maturation of the telecom industry. Many of the fundamental challenges facing digital optical computing remain difficult to achieve simultaneously in a simple device.

For this reason, many attempts to leverage the capabilities of optics avoid a digital electronic computing paradigm altogether and instead target specialized tasks, including A/D conversion [41], amoeba-inspired processing in quantum dots [42], and reservoir computing [43], [44]. The utility of an overspecialized optical "hardware accelerator" or "coprocessor" has so far been outweighed by the cost of commercial platform development, although many unconventional approaches succeed in exploring new and interesting intersections of computing and physics [45]. The proposed architecture avoids overspecialization with its many configuration freedoms—both in design layout and in field-tunable interconnection parameters. A particular

interconnect configuration, which determines the behavior and function of a distributed processing system, is very different than procedural program, where operations are represented by a stack of instructions interpretable by a Turing machine.

This absence of procedural programmability is a challenge for the analysis and design of all neuron-inspired architectures, but also one of their biggest advantages. Processors that relinquish a framework of immutable execution could exhibit enhanced aptitude to self-organize and adapt to uncertain environments without programmer input [46], [47]. We believe that the architecture proposed here exhibits important properties of a computing system potentially capable of sophisticated and widely applicable large-scale information processing (see Section IV-B), but classify it as a "scalable photonic processor" to emphasize the fact that it does not pursue a symbolic instruction model of general purpose computation. Among unconventional optical processing paradigms, neural networking is perhaps the most commonly examined class of models.

C. Optical Neural Networks

Optical technologies for interconnection have long been recognized as potential media for artificial neural network architectures, which rely on parallel communication performance as much as—if not more than—parallel operation of computational gates (a.k.a. neurons). Attempts to realize the throughput, dissipation, and cross-talk advantages of optics in a neurocomputing context, while promising in many cases, have so far encountered barriers in reliability, scalability, and cost. A review of optical neural networks (ONNs) is contained in [1].

For the most part, approaches to ONN interconnection have focused on spatial multiplexing techniques, including configurable spatial light modulation [48], matrix grating holograms [49], and volume holograms [50], [51]. Although they are dense techniques for all-to-all interconnection, free-space and holographic devices are difficult to integrate and also require precise alignment. Systems that are non-integrable or that require exotic integration processes have extreme difficulty matching CMOS systems in cost or practical scalability.

Coherent interference effects in many-to-one coupling [52] are particularly relevant to neural networks with large fan-in. Phase-sensitive designs of spiking optical neurons, such as [26], [27], must introduce methods to control the relative phases of signals originating from distinct computational primitives. Semiconductor optical devices that implement a Hopfield (non-spiking) model have used WDM to avoid mutual interference [48], [53]. Using WDM as a non-spatial multiplexing technique, broadcast-and-weight is compatible with commercial PIC integration and can exploit this spatial indeterminism to bestow a distributed architecture with structural features not possible with holographic or free-space systems, as discussed in Section IV-A.

III. System Architecture

The proposed architecture for photonic spike processing consists of three aspects: a protocol, a node that abides by that protocol, and a network medium that supports multiple

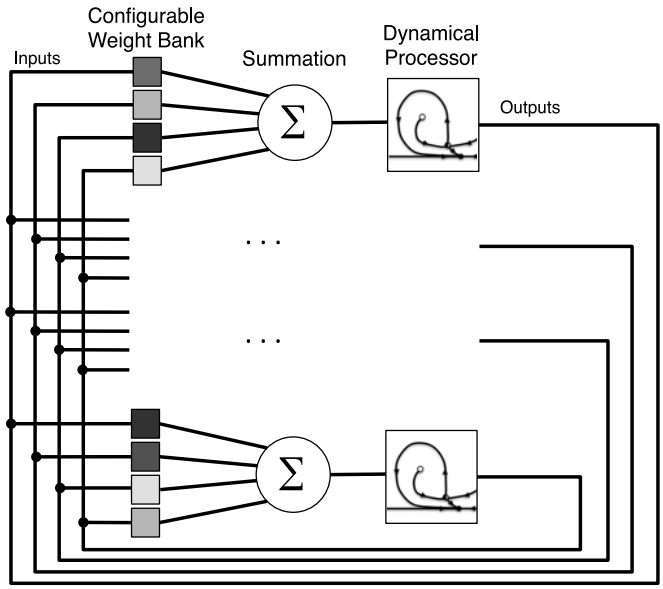

Fig. 1. Functional model of a spiking neural network, depicting four neurons. Each neuron has one output signal, which is sent to multiple other neurons. Input signals are independently weighted by an analog coefficient (represented by grayscale value) before summation. The summed signal drives a dynamical processing model, such as spiking leaky integrate-and-fire (represented by the phase portrait of an excitable system).

connections between these units. Broadcast-and-weight is a WDM protocol in which many signals can coexist in a single waveguide and all participant units have access to all the signals. The processing-network node (PNN) is a primitive unit that performs the physical and logical functions required for broadcast-and-weight networking and neuromorphic processing, respectively. The broadcast loop (BL) defines the medium in which a broadcast network exists and physically links a group of PNNs to one another. Although the authors have made every attempt to present these aspects in a linear fashion, they are logically intertwined; a more thorough discussion of design justifications is deferred until after the aspects are presented together.

In every neural network model, each node receives signals from many other nodes, performs some process, and transmits copies of a single output signal to multiple receiver neurons (see Fig. 1). Each input is modulated independently by a constant multiplier (a.k.a. weight), which can be positive, negative, or zero. After weighting, all inputs to the neuron are summed, before modulating a nonlinear dynamical element: in this case, a laser neuron device. The configuration of the system is determined by its weight matrix, where element w_{ij} signifies the strength of the connection from neuron i to neuron j. A single transmission device can not alter the polarity of a signals represented as optical power, so effective neural weighting requires two optical filters per channel dropping power into a balanced push-pull photodetector in order to implement both positive and negative weights. A processor can exhibit a large variety of behaviors through reconfiguration of the weight matrix, although this weight tuning happens on timescales much slower than spiking dynamics. The problem of neural networking

contains prominent one-to-many (multicast) and many-to-one (fan-in) components. In the case of spiking networks, communication signals are pulses: binary in amplitude and asynchronous in time. For interconnecting signals with spikes represented as physical pulses (as opposed to digital packets as in AER), temporal multiplexing and switch-based routing techniques are not viable strategies because spike timing is an informatic dimension unavailable for multiplexing. The goal of our network design will be to support a large number of parallel, asynchronous, and reconfigurable connections between a distributed group of photonic processing primitives that is compatible with the approach of spikes represented physically as optical pulses.

A. Broadcast-and-Weight

WDM channelization of the spectrum is one way to efficiently use the full capacity of a waveguide, which can have useable transmission windows up to 50 nm wide (>1 THz bandwidth) [54]. In fiber communication networks, a WDM protocol called broadcast-and-select can create many potential connections between nodes: the active connection is selected, not by altering the intervening medium, but rather by tuning a filter at the receiver to drop the desired wavelength [55]. We present a similar protocol for a spike processing network and call it "broadcast-and-weight." It differs by allowing multiple inputs to be dropped simultaneously and with intermediate strengths between 0% and 100%.

Broadcast-and-weight consists of a group of nodes sharing a common medium in which the output of every node is assigned a unique transmission wavelength and made available to every other node (see Fig. 2). Each node has a tunable spectral filter bank at its front-end. By tuning continuously between 0–100% drop states, each filter drops a portion of its corresponding wavelength channel, thereby applying a coefficient of transmission analogous to a neural weight. The filters of a given receiver operate in parallel, allowing it to receive multiple inputs simultaneously. An interconnectivity pattern is determined by the local states of filters and not a state of the transmission medium between nodes. Routing in this network is transparent, parallel, and switchless, making it ideal to support asynchronous signals of a neural character.

The ability to control each connection, each weight, independently is critical for creating differentiation amongst the processing elements. A great variety of possible weight profiles allows a group of functionally similar units to compute a tremendous variety of functions despite sharing a common set of available input signals. Reconfiguration of the filters' drop states, corresponding to weight adaptation or learning, intentionally occurs on timescales much slower (μs or ms) than spike signaling (ps). A reconfigurable filter could, for example, be implemented by a microring resonator whose resonance is tuned thermally or electronically. In a group of N nodes with N wavelengths, each node needs a dedicated weighting filter for all $(N-1)$ possible inputs plus one filter at its own wavelength to add its output to the broadcast medium. The total number of filters in the system would thus scale quadratically with N^2. A filter design example is given in Section III-D.

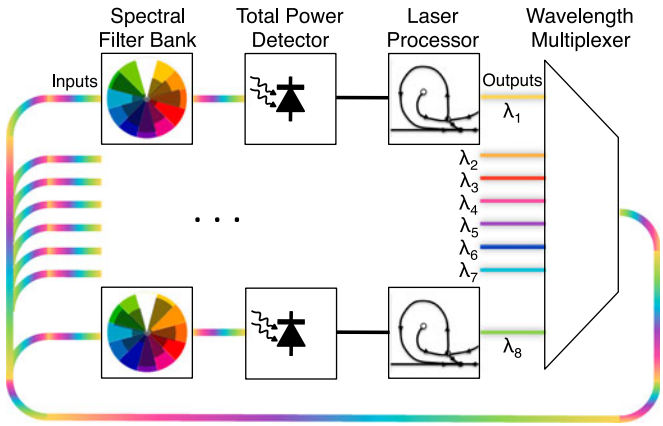

Fig. 2. Optical broadcast-and-weight network showing parallels with the neural network model of Fig. 1. An array of source lasers outputs distinct wavelengths (represented by solid color). These channels are wavelength multiplexed (WDM) in a single waveguide (multicolor). Independent weighting functions are realized by spectral filters (represented by gray color wheel masks) at the input of each unit. Demultiplexing does not occur in the network. Instead, the total optical power of each spectrally weighted signal is detected, yielding the sum of the input channels. The electronic signal directly drives a laser processing device, such as the excitable laser proposed in [32].

B. Processing-Network Node

In a biological neural network, the complicated structure of physical wires (i.e., axons) connecting neurons largely determines the network interconnectivity pattern, so the role of neurons is predominantly computational (weighted addition, integration, thresholding). The contrasting all-to-all nature of optical broadcast saddles the photonic neuron primitive units with additional responsibilities of network control (routing, wavelength conversion, WDM signal generation, etc.).

The proposed design of a PNN can perform all of these necessary functions, achieving compactness by flattening the dual roles of processing and networking into a single set of devices. It attains rich computational capabilities by leveraging analog physics offered by optoelectronics. Overall, the PNN is an unconventional repurposing of conventional optoelectronic devices, thereby appearing as a strikingly simple circuit with potential to generalize to existing—and prospective—photonic platforms. One possible implementation of a PNN is depicted in Fig. 3, while the dual purpose of its devices are summarized in Table I.

The PNN interacts with a WDM waveguide via two tunable filter banks. One filter bank represents the weights of excitatory (positive) input connections while the other controls inhibitory (negative) inputs. These weight profiles could be stored in local co-integrated or off-chip CMOS memory. The two weighted (i.e., spectrally filtered) subsets of the broadcast channels are dropped—without demultiplexing—to a balanced photodiode pair. Photodetectors output a current that represents total optical power, thus computing the weighted sum of WDM inputs in the process of transducing them to an electronic signal, which is capable of modulating a laser device. The balanced photodiode configuration enables inhibitory weighting, which is an essential capability of any neural network.

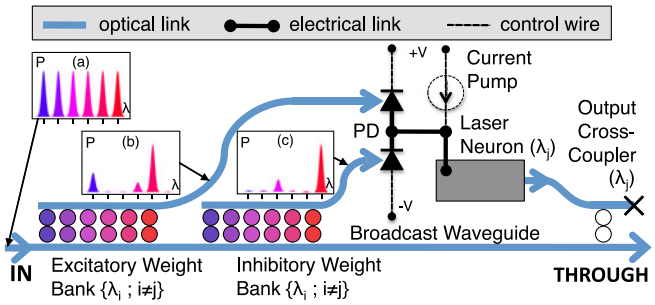

Fig. 3. Processing-network node (PNN) coupled to a broadcast waveguide. The front-end consists of two banks of continuously tunable microring drop filters that partially drop WDM channels that are present. Two waveguide integrated photodetectors (PDs) convert the optical signal to an electronic current and perform summation operations on the weighted excitatory and inhibitory inputs. A short wire subtracts these photocurrents and modulates current injection into an excitable laser neuron, which performs threshold detection and pulse formation in an optical cavity. The output of the laser is coupled back into the broadcast waveguide and sent to other PNNs. Insets represent example spectrograms of the waveguides. (a) Broadcast waveguide with 6 WDM channels: (b) three of these channels are shown partially dropped into the excitatory PD, and (c) two other channels are shown partially dropped into the inhibitory PD. The channel subsets that are dropped are determined by the tuning state of each filter (driving circuitry not shown).

TABLE I
CORRESPONDENCE BETWEEN COMPUTING AND NETWORKING FUNCTIONS IN THE PRIMARY SIGNAL PATHWAY

Element	Process Function	Network Function
Adaptive filter bank	Weight multiplication	WDM drop-and-continue
Photodetector	Addition/subtraction	Multiwavelength fan-in
Gain medium	Temporal integration	Laser modulation
Excitable laser	Threshold detection	Clean pulse generation
Output coupler		WDM add

Total optical power detection of a still multiplexed signal is a relatively rare technique because it irreversibly strips WDM signals of any trace of their identifying wavelength. This property has been exploited in several applications including subcarrier optical multiplexing [56], a multi-input OR function [57], and analog RF photonic signal processing [58]; nevertheless, it is counterproductive in the majority of situations. Information about a signal's origin is desirable in multiwavelength communication systems and is maintained by demultiplexing prior to detection. In the neurocomputing context however, this destruction of channel information is precisely correspondent with the summation function. A photodiode can therefore be viewed in this sense of dual purpose, not just as a transducer, but also as an additive computational element capable of many-to-one wavelength fan-in.

The PNN front-end is not subject to well-known optical-electronic-optical (O/E/O) conversion overhead. The cost, energy, and complexity typically involved in O/E/O are due not, in fact, to the physical transduction itself but instead to the electronic receiver stages (i.e., amplification, sampling, and quantization) that normally follow detection in fiber communication links [40]. The "receiver-less" pathway connecting photodiodes to laser neuron is not significantly affected by dispersion or electromagnetic interference (EMI) in this case because it can be made very short ($\sim 20\,\mu$m) regardless of fan-in degree.

The electronic signal from the balanced photodetector pair modulates a laser processor, which performs some dynamical and strongly nonlinear process, described in more detail in [31], [32]. The modulated laser gain medium is an active optical semiconductor, which acts as a subthreshold temporal integrator with timeconstant equal to carrier recombination lifetime. The laser system itself acts as a threshold detector, rapidly dumping energy stored in the gain medium into the optical mode when the net gain of the cavity crosses unity, much like a passively Q-switched laser biased below threshold. In this way, it emulates one of the most critical dynamical properties of a spiking neuron—excitability—on picosecond timescales. Although the possibility of WDM was not explicitly discussed in prior work, the lasing wavelengths of an array of excitable distributed feedback lasers could be tailored by altering the pitch of their gratings [59].

By generating clean, stereotyped pulses at a single wavelength, the laser provides the optical signal necessary for broadcast-and-weight networking. All light can be generated and detected on-chip. In addition, excitable lasers effectively provide gain, since large pulse responses can be triggered by weak input pulses. If excitable gain is sufficient to counteract insertion and fan-out losses, this means that, in principle, active optoelectronics would not be necessary outside of the PNN module.

Finally, an output coupler adds the generated signal to the broadcast waveguide. Other wavelengths are nominally unaffected by this coupler, but any incoming signals at the PNN's assigned wavelength will be completely dropped and terminated, avoiding collision with the newly generated output.

C. Broadcast Loop

The final aspect of the proposed networking architecture is the physical medium that transports WDM optical signals between the output couplers and input spectral filter banks of a group of PNNs. Since routing is already performed by the PNN filters, the broadcast medium must simply implement an all-to-all interconnection, supporting all N^2 potential—not necessarily actual—connections between participating units. This role can be performed by a single integrated waveguide with ring topology, which we refer to as a BL. A broadcast-and-weight cell thus consists of several PNN primitives coupled to a BL medium, as illustrated in Fig. 4. Its ring shape is reminiscent of metropolitan fiber networks, though the neuromorphic processing implications of the BL are worthy of further consideration.

The BL waveguide is fully multiplexed at all points along its length. Most signal power is allowed to continue through a PNN, even if a portion of it is dropped. This technique called drop-and-continue is an instance of lightpath splitting, where the information carried by an optical channel can be copied passively and instantaneously, albeit with a reduction in power [60]. The weight-dependent signal power distribution of drop-and-continue does create an undesirable interdependency between

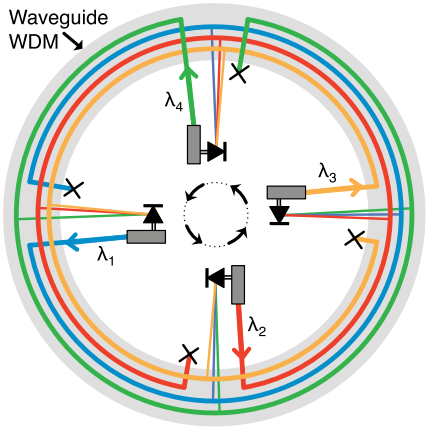

Fig. 4. Conceptual diagram of a broadcast loop (BL). The loop waveguide carries WDM channels from all participating PNNs, so each PNN can detect a configurable subset of all channels. The PNN laser then outputs its signal, a function of those inputs, on its unique wavelength channel. Once a signal transverses the BL, it is completely dropped and terminated by its originating unit to avoid interference between different parts of a channel. Filter banks and inhibitory pathways not shown.

filter weights at different neurons, which could present a control problem in adaptive systems. Drop-and-continue is a physical solution to optical multicasting that can radically reduce network traffic for a given virtual interconnect density [34]. In the BL, this technique reaches it's maximum potential, supporting N^2 independent interconnections in a waveguide with only N channels.

An example of a folded layout for tight packing is shown in Fig. 5. Multiple BLs integrated on the same chip could interact by simply designating interfacial PNNs: nodes that receive inputs from one BL and transmit into another (bottom of Fig. 5). In this way, a unified processing system consisting of multiple BLs can be created without any additional arbitration, routing, or device technology. BLs interacting via interfacial PNNs constitute distinct broadcast media and can thus reuse the same optical spectrum, much like a cellular telephone network reuses spectrum geographically. Unlike a cellular phone network however, the operation of these broadcast media is dissociated from their exact geometry, as long as the loop topology is present. The associated spatial freedoms will be seen to yield a promising variety of multi-BL architectures (see Section IV-A).

D. Design Example

The design of tunable filter banks for WDM weighting can proceed similarly to that of wavelength demultiplexers based on microring resonators (MRRs) in conventional digital interconnects. In [54], the FSR-limited maximum wavelength count for a silicon WDM link was found to be $N = 62$ for a transmission window of 50 nm and channel spacing of $\Delta\lambda_0 = 5.3$ linewidths (0.8 nm). Heterogeneous integration platforms incorporating III/V active sections and passive silicon-on-insulator (SOI) waveguides have demonstrated broadband photodetector responsivity and, with proper design, single-mode lasing over a 45 nm band (1525–1570 nm) [61]. The transfer function of a resonator drop filter is approximated by a Lorentzian function:

$$T(\delta) = \frac{1}{1+\delta^2}, \text{ where } \quad \delta = \frac{Q}{\omega_0}(\omega - \omega_0) \tag{1}$$

in which δ is linewidth-normalized frequency, Q is quality factor ($Q \approx 10,300$), and ω_0 is peak center frequency. The extinction ratio of a single filter is $R = T_{\max}[dB] - T_{\min}[dB] = T(0)[dB] - T(\omega_{\text{tun}})[dB]$, where ω_{tun} is the maximum tuning range in linewidths. For broadcast-and-weight, analysis of cross-talk must also take into account that the resonant frequency of each filter is shifted in order to control the weight applied to its channel. The worst case cross-talk X_{ij} (defined as in [54]) is not identical between upper and lower neighbors because the filter resonance moves towards longer wavelength channels when detuned from center. For the jth neighbor: $X_{j0} = T(j\Delta\omega)[dB] - T(0)[dB]$ and $X_{0j} = T(\omega_{\text{tun}} - j\Delta\omega)[dB] - T(0)[dB]$, where $\Delta\omega$ is channel spacing. Insertion loss on the i^{th} channel is $I = \left(1 - R^{-1}\right) \cdot \Pi_{j<i}\left(1 - X_{ij}\right) \cdot \Pi_{j>i}\left(1 - X_{ji}\right)$ where I, R, and X are in linear units.

For a specified tolerable performance of $R > 13$ dB, $\{X_{j0}, X_{0j}\} < -13$ dB, and $I < 0.35$ dB, we find WDM parameters of $\omega_{\text{tun}} = 4.4$ (0.66 nm) and $\Delta\omega = 8.8$ (1.3 nm) meet this specification. With the 45 nm gain band of hybrid III-V/SOI lasers, these parameters lead to a channel number $N = 34$ per BL. The approximate footprint of a single filter bank in this case is $34 \times 16 = 540$ μm^2, compared to ~4,000 μm^2 for the active devices in a single PNN. The corresponding BL footprint is $34 \times 4{,}540 = 0.15$ mm^2. The BL waveguide must be at least $34^2 \times 4\,\mu\text{m} = 4.6$ mm long to physically accommodate this number of filters, contributing a minimum power penalty of about 0.4 dB, given SOI waveguide loss [62]. We have made the simplifying assumptions that every connection has a dedicated tunable MRR filter, these filters are all critically coupled to the bus waveguide, and that they are single-pole (i.e., single-MRR). Further investigations that depart from these simplifying assumptions could likely improve performance and maximum number of channels (for example, using double-pole MRRs for steeper filter rolloff [54]).

Power budget is also a very important design consideration; however, the analysis of noise and signal power in conventional digital interconnects can not be mapped trivially to the present system. Although similar noise mechanisms are present (e.g., ASE, cross-talk, etc.), the relationship between SNR and spike error rate in an optical spiking link requires further investigation. For a full system design, the tolerance of overall system function for communication errors must also be specified. This tolerance is application-dependent, but likely relaxed compared to digital systems, due to the statistical and intrinsically noisy nature of neuromorphic algorithms.

IV. Discussion

The broadcast-and-weight protocol is a novel approach for combining neuromorphic processing and optical networking, based on deep-seated correspondences between the chosen models of processing and networking. This combination gives rise

to novel properties that are native neither to optics nor to neuroscience. Optical WDM in a waveguide gives the architecture special spatial freedoms, which are not observed in other hardware neuromorphic systems. These freedoms will be discussed with respect to their practical consequences to layout and organizational flexibility. The spiking paradigm has reciprocal effects on optics as an information processing substrate. We find that many of the common challenges of robustness, cascadability, and scalability faced by conventional optical logic architectures can be addressed, a possibility largely attributable to the unconventional paradigm.

A. Multi-BL System Layout and Organization

A means to interface different BLs was initially introduced in Fig. 5 to reuse spectrum on-chip. Although PNNs in different loops can interact indirectly via interfacial PNNs, a multi-BL system does not exhibit the same all-to-all potential interconnection observed in a single BL. This could cause informatic fragmentation and bottlenecks between different parts of a system with many interfaced BLs, effectively neutralizing the computational usefulness of scaling the node count. We argue that interconnect sparsity resulting from spectral reuse is not necessarily detrimental to overall computational complexity, provided design can follow appropriate principles. When determining structural constraints in distributed processing networks, communication and computation become fundamentally intertwined, so design rules for organizing multi-BL architectures must shift to invoke concepts outside of the field of communication networks. We find that the ability to incorporate these distributed processing principles in an optical system is made possible by a special topological property of broadcast-and-weight, which we call spatial layout freedom.

1) Spatial Layout Freedom: A BL waveguide can manifest arbitrary shape in order to accommodate any layout of a group of PNNs; this stipulation contrasts nearly all other approaches to physical neuromorphic architectures (e.g., cross bar arrays or holographic matrix-vector multipliers), where the layout of computational primitives follows from the particular parallel networking approach. In a situation where signals are distinguished based on their position, wire, or wavevector, physical layout inherits the geometrical constraints of the interconnect, which can give rise to tangible limitations to interconnect structure (e.g., Rent's rule [63]). Biology can avoid multiplexing altogether by using dedicated wires (i.e., axons) for every connection. However, this 3-D approach is not possible with state-of-the-art quasi-2-D fabrication techniques. While the exact implications of this dimensional disparity are beyond the current scope, one can assert provisionally that any conservation of spatial degrees of freedom could be supremely important in neuromorphic engineering.

In the broadcast-and-select protocol, spatial degrees of freedom are essentially undetermined: node identity is distinguishable based on wavelength alone. In addition, the large bandwidth-distance product of optical waveguides means the corrupting role of dispersion remains small over a range of spatial scales, compared to electrical transmission lines [64].

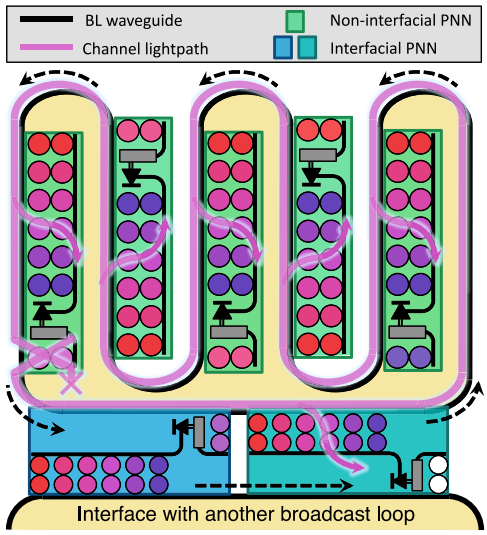

Fig. 5. Example folded layout of a broadcast-and-weight cell showing 5 PNNs (delimited by green areas) and two interfacial PNNs (blue areas) coupled to a contained BL (tan area). The lightpath of one channel (magenta) is shown traversing the BL waveguide and branching into multiple filter banks. Originating and terminating in the leftmost PNN, this signal can be partially dropped into any of the PNNs around the BL. Each processing node must transmit on a unique wavelength channel, except the outgoing interfacial node (lower right), which transmits into a different loop. Each node's filter bank drops a linear superposition of the present channels, except the incoming interfacial node (lower left), whose inputs are derived from another loop. Inhibitory pathways not shown.

Although WDM and bandwidth-distance properties of optics have been used for decades in communication networks, distributed processing consequences of spatial indeterminacy have not been explored. This is not a matter of oversight, but rather context. Fiber telecom networks transport signals between geographic locations, a purpose intrinsically tied to space. On the other hand, processing networks transport signals between a group of computational nodes; it makes no essential difference where its nodes or its signals are located. At any spatial scale, BL implementation relies on an identical device repertoire (i.e., filters, photodetectors, and excitable lasers), with the exception perhaps of bus waveguide amplifiers that are needed to counter distance dependent loss in a silicon waveguide. Spatial invariance in multiplexing protocol, signal transmission, and device technology—in the context of distributed processing—results in the possibility to implement interesting and important structures in multi-BL architectures.

Fig. 6 illustrates a multi-BL structure, demonstrating key features of hierarchical organization. Each BL reuses the same spectrum and WDM channelization, but can represent different hierarchical levels of organization. A level-1 BL interfaces with other level-1 BLs (via "lateral" PNNs) and a level-2 BL (via "uplink" and "downlink" PNNs). Interfacial PNNs can be thought of as regular PNNs whose input spectral weight bank receives the broadcast signals of a different BL (Fig. 5). While similar in some ways to routing interfaces in conventional optical communication networks (which can also have hierarchical organizations), the PNN interfaces are spike processors that intrinsically transform information while transporting it. As a

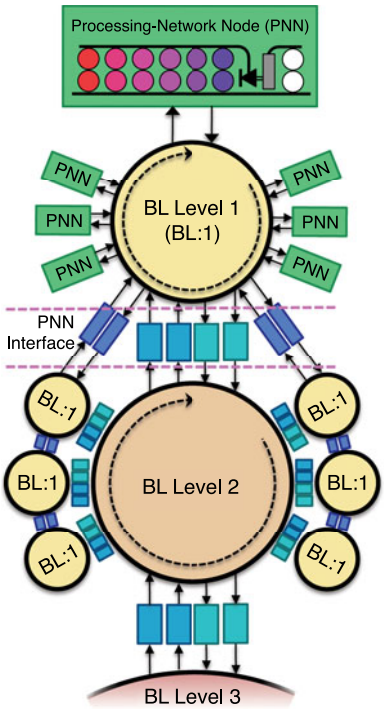

Fig. 6. Hierarchical organization of the waveguide broadcast architecture showing a scalable modular structure. Colored rectangles represent PNNs. Green PNNs indicate input and output coupling to the same broadcast loop. Blue PNNs interface between distinct BLs and are classified as "uplink," "downlink," or "lateral" varieties based on their position in the hierarchy. Each transmitting PNN has a unique output wavelength within its given broadcast space, but spectrum is reused between different BLs.

result of the processing done in PNN interfaces, network nodes in a given BL can not directly send their outputs to nodes in other BLs, and multi-BL systems can no longer implement all-to-all interconnects. Instead of attempting to faithfully transfer any one signal from one BL to another, the PNN interfaces create mutual informatic relationships that extend beyond BL boundaries. At the same time, PNN interfaces do not experience additional buffering or wavelength allocation constraints, and the BL communication load is constant across different levels of the hierarchy instead of growing exponentially as in pure communication networks.

Fig. 7 shows a layout that corresponds to the network diagram of Fig. 6. The lowest level is a tightly packed group of computational primitives connected by a folded loop (see Fig. 7(c)). Some computational primitives can interface with other loops, either directly with nearby first level loops, or with a second level loop that connects physically distant components on the chip-scale. The second level loop (see Fig. 7(d)) has a similar functionality compared to the first level, but it occupies a much larger area and represents a more complex dynamical processing network. Although the chip scale corresponds to just the second level in this example, intermediate levels on chip are entirely possible. Continuing in this direction of hierarchical levels, a multi-chip system based on fiber loops (see Fig. 7(e)) could be considered. Interfacing multiple optoelectronic chips all-optically through a transparent fiber-waveguide path represents an interesting possibility for further investigation.

Spatial layout freedom can be viewed as a powerful tool to combat the sparse interconnection constraints inherent in multi-BL spectral reuse and allow a wide potential variety of system organization. However, determining particular multi-BL organizations and the number of PNNs allocated at each interface represent significant design challenges. Design parameters that impact network structure fundamentally exceed pure communication theory and must invoke theories of distributed computation, such as complex network topology and cortical organization.

2) Organization Principles for Multi-BL Architectures: Developments in complex network theory have recently been applied to understand aspects of structure, organization, and collective dynamics in cortical networks [65], and insights from this field could be used to guide multi-BL system design. Complex network theory describes relationships between interconnection patterns (i.e., graph topology) and dynamic functionality in distributed systems, which contrasts with the study of information capacity in static states or isolated communication channels. While the goal of neuron-inspired processing should not be perfect emulations of biological networks, the study of cortical connectomics (i.e., biological neural network structure) also provides examples of the types of topological features that may be relevant for processing tasks in neuron-inspired systems. Important aspects can be judged with the tools of complex network science and connectomics, which enable the abstraction of relevant metrics of informatic and computational complexity in distributed systems.

For example, a complex network metric called "small-worldness" describes some networks that lie between an ordered and random interconnectivity pattern. "Small-worldness" is engendered by both high clustering coefficient (i.e., cliquishness) and short average path length (i.e., sparse long-range connections) [66]. In complex systems, small-world networks have been associated with dynamical complexity [67] and information integration over multiple spatial scales [68]. Small-world characteristics are also observed in anatomical networks, ranging from the simplest animal nervous system (*C. Elegans*), to mammalian cortex, which has a consistently modular and hierarchical organization throughout [69].

These biological and mathematical insights could provide evidence to guide organizational design principles of neuromorphic processing systems. Spatial layout freedom means a BL can fully interconnect a tightly packed group of processing nodes, or it can run over an entire chip area. This coexistence of large fan-in and long-range connections is a physical correlate of the simultaneous clustering and short path lengths that typify small-world networks.

In order to realize small-world topological properties in an artificial neural network, an interconnect implementation must support connections over a range of spatial scales. Electrical wires exhibit a bandwidth-distance-energy tradeoff that impedes this goal [64]. Systems based on spatial multiplexing in holograms or cross-bar arrays cannot be easily detached from a characteristic length (e.g., diffraction length) and have very

little flexibility or potential to scale hierarchically. Spatial layout freedom as described above could grant the flexibility required to meet these goals, making broadcast-and-weight architectures uniquely suited, among artificial hardware systems, to explore computationally efficient and biologically-relevant network topologies.

Based on qualitative similarities between organizational abilities of multi-BL systems and principles of complex and cortical networks, we have hypothesized that the proposed architecture is capable of enacting salient processing structures. Further inquiry into multi-BL architecture design must incorporate principles of complex network theory, likely including, but not limited to, the idea of small-worldness. Concretization of the corresponding design rules represents a formidable research problem, which lies in the intersection of linear lightwave networks and complex system science.

B. Feasibility of Photonic Processing With Spikes

In this section, we will briefly consider how three aspects of practical feasibility (cascadability, robustness, and scalability) in photonic processing are impacted by adopting a spike processing paradigm. Cascadability is the ability of a computational element to drive multiple stages of similar devices with fidelity in the presence of noise. Robustness refers to a system's potential to mitigate the effect of device defects—inevitable in large-scale integration—on overall functionality. Scalability is an architecture's capacity to increase in size and complexity, which requires a system format able to accommodate modular expansion without performance degradation.

1) Cascadability: In digital electronic design, a logic gate needs power gain to fan-out to multiple other gates, and it must have logic-level restoring behavior to suppress noise. These conditions usually imply cascadability in electronics, yet a more multifaceted notion of cascadability applies to an optical device due to the extra dimension of wavelength (or phase). This extra degree of freedom can be a major boon to functionality in an optical system (e.g., WDM) but can introduce vulnerabilities to new sources of uncertainty (e.g., wavelength drift). In particular, systems that exploit WDM can suffer from a need for wavelength conversion.

The proposed PNN co-integrates the complementary physics of optics and analog electronics in order to address cascadability issues in WDM. The PNN curtails propagation of phase/wavelength noise from one stage to another by interleaving optical representations with an analog electronic part of the primary signal pathway. The photodiode-laser setup "converts" information from multiwavelength inputs onto a single wavelength output, physically capable of driving other PNNs. However, total power detection for wavelength fan-in is inseparable from an analog summation function. While this effect would corrupt channel information in digital signals, the summation is precisely correspondent with weighed summation in models of neuromorphic processing.

All-or-nothing output quantization is critical in spiking paradigms because the significant amount of analog processing is vulnerable to amplitude noise. The excitable laser employs cavity-mediated optoelectronic interactions to realize spiking dynamics at ultrafast timescales, which allow it to perform hybrid analog-digital information transformations in a small footprint [31]. These dynamics, shared by spiking biological and analog CMOS neurons, prevent noise generated in analog portions of the pathway from propagating through the system and eventually corrupting signal integrity. Fan-out can pose a problem to optical processors because splitting is accompanied by an N-fold reduction in signal power. This loss could be counterbalanced by laser excitable gain, in that small input pulses can trigger the release of a much larger quantity of stored energy, or with additional waveguide amplifiers in the BL.

Spikes carry information predominantly in their timing, so time skew has the potential to corrupt signals. The authors of [70] noted that differences in electronic and optical signal transmission can cause timing requirements that make the leap from combinatorial logic to sequential logic highly nontrivial. However, since synchrony is not a critical aspect of the spike

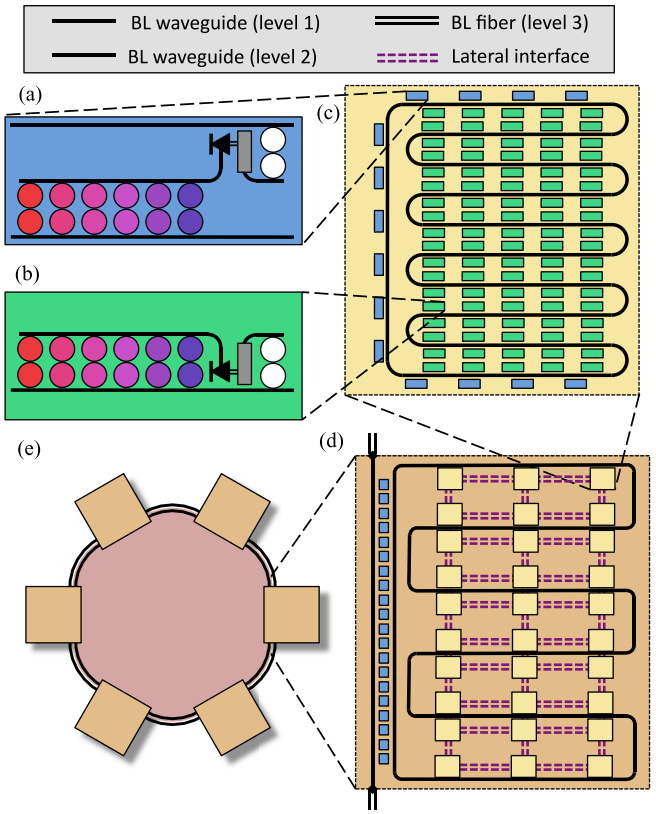

Fig. 7. An example layout strategy for a hierarchical network demonstrating the scale-independent nature of a waveguide BL. Computational primitives are classified as (a) interfacial PNN, whose output is coupled into a different BL waveguide than its inputs and (b) non-interfacial PNN, which transmits and receives in the same BL. (c) A broadcast-and-weight network constitutes the first-level of hierarchy and consists of a group of potentially all-to-all connected computational primitives. In this case, it takes a folded shape for the sake of packing efficiency. (d) A chip-scale second-level broadcast network interconnects the interfacial PNNs from many first-level BLs. First-level BLs can also interface directly via lateral interfacial PNNs (purple dotted lines). (e) A multichip third level network illustrating a compatibility with fiber implementations of a BL. The broadcast-and-weight network is conceptually the same as in other levels, but the BL waveguide consists of coupled fibers and integrated waveguides.

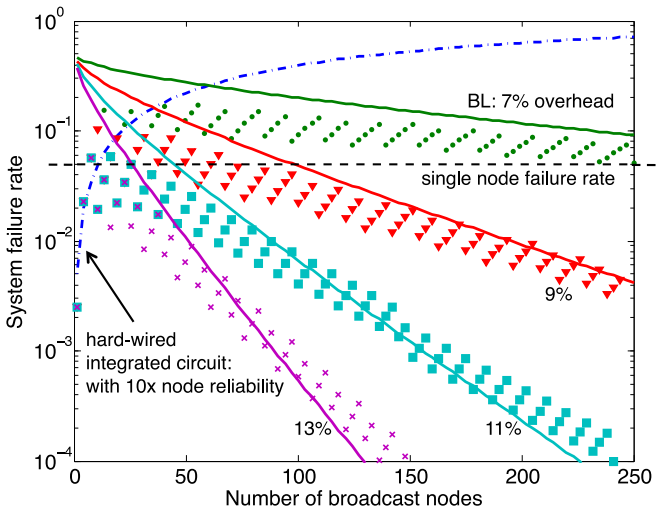

Fig. 8. System failure rate as a function of the number of nodes comparing a conventional hard-wired circuit (blue dash-dot line, Equation (2)) to broadcast-and-weight systems with varying amounts of hardware overhead (7%: green circle, 9%: red triangle, 11%: cyan square, and 13%: magenta cross). The exact failure rate of the BLs (markers, Equation (6)) differ from the approximate error function curves (solid lines, Equation (4)) due largely to integer rounding. Even though the hard-wired system is shown here with nodes of 10 times higher reliability ($5 \cdot 10^{-3}$ versus $5 \cdot 10^{-2}$ failure rate), the systemic reliability of a BL can be much greater than the hard-wired system and even the reliability of a single element (black dotted line).

processing paradigm, strict conformity of timing parameters is not necessary. The asynchronous nature of broadcast-and-weight provides a mechanism to perhaps even exploit heterogeneity in spike timings in order to implement advanced spatiotemporal algorithms, such as [71]. On the other hand, the effect of noise on pulse timing (i.e., jitter) is relevant in determining spike precision and channel capacity.

2) Robustness: Suppose a given distributed processing task requires n computational primitives. Each device has some fixed reliability: the probability that it will work successfully p_{succ}. Since the system requires *all* devices working, its failure rate is given by

$$P_{\text{fail}} = 1 - p_{\text{succ}}^{n}. \tag{2}$$

Systemic failure rapidly approaches certainty as the system size (i.e., node count) increases. This unreliability is particularly important for large-scale integrated systems since, a defective transistor or laser device cannot simply be replaced after the fact. Robustness can be improved by increasing device yield, a strategy that is not always practicable, or by incorporating hardware redundancy called overhead. It is impossible to know ahead of time which devices will fail, so overhead must cover every possible failure, even though each is highly unlikely. If each primary device is given a backup device (100% overhead), the majority of overhead hardware will remain unused, and a joint failure of both primary and backup devices could still disable the system. More sophisticated ways of incorporating redundancy based on coding theory can be applied in special cases, but no general code theoretic approach to robustness in Boolean systems has yet been identified [72].

The broadcast-and-weight network can easily incorporate small amounts of hardware overhead. Since all PNNs have access to all signals in a single BL, they can be swapped interchangeably in the event of device defect or death. The PNNs are functionally similar, so any unused PNN can virtually swap its interconnection relationships with any defective PNN by exchanging filter bank weights. Overhead PNNs therefore do not backup a single primary PNN, but rather cover all possible failures in the BL. Virtual swapping through reconfiguration can react to specific failures that occur both during fabrication or in the field. Programming a reconfiguration to avoid defects can be very energy and computation hungry in some systems (i.e., field-programmable gate arrays) due to the intensive problems of placement and routing associated with mesh networks [73]. In contrast, a broadcast network has no corresponding constraint in mapping automata to devices, trivializing the hardware optimization problem.

The ability to easily swap the role of every hardware primitive means that system success now requires *any* n processors to work out of a total of $m = \lceil (1+a)n \rceil$ PNNs in the BL, where a is overhead ratio. If the number of working PNNs $k \in (0, 1, \ldots, m)$ is a Poisson random variable

$$P\left[k\right] = \binom{m}{k} p_{\text{succ}}^{k} \left(1 - p_{\text{succ}}\right)^{m-k} \tag{3}$$

$$P_{\text{fail}} = \sum_{k=0}^{n-1} P[k]. \tag{4}$$

For large values of n, this failure rate can be approximated

$$P(k) \approx \text{Norm}\left(k\,;\, mp_{\text{succ}}, m(1-p_{\text{succ}})\right) \tag{5}$$

$$P_{\text{fail}} \approx \frac{1}{2}\text{erfc}\left(\frac{mp_{\text{succ}} - n}{\sqrt{2m(1-p_{\text{succ}})}}\right) \tag{6}$$

where Norm $(k; \mu, \sigma^2)$ is a Gaussian function with mean μ and variance σ^2 and erfc $(\cdot)$ is the complementary error function. System failure rate as a function of network size is plotted in Fig. 8, comparing the robustness of hard-wired systems to broadcast-and-weight systems with varying amounts of hardware overhead. The system with swappable nodes inverts the conventional trend, exhibiting a failure rate that decreases exponentially with the nominal node count. Surprisingly, systemic reliability can even be better (in some cases by orders of magnitude) than the reliability of a single node.

This mechanism of robustness through swapping could be very useful in other on-chip photonic networks; however, it does not extend arbitrarily to computational models outside of neuromorphic processing. Only processing elements invariant to input ordering (e.g., addition, NAND, etc.) allow for swapping of nodes. In most other processing models (e.g., Fredkin gates, CPU cores, etc.), the sequence of different inputs must remain distinguishable to a processor. This invariance to input sequence in a summation corresponds to a photodetector destroying wavelength information, which is a key compatibility between the photonic physics and neuromorphic function of the PNN.

3) Scalability: Any processing technology is also subject to notions of scalability since it will be compared to the highly developed and continually advancing microelectronic standard. Broadcast-and-weight can expand to multi-BL architectures due to modular abstractions of the PNN and BL, in which performance-limiting electrical links are very short and memory of weight values can be locally co-integrated. Modern trends in photonic integration practices that support WDM techniques could grant photonic spike processing architectures a pathway to low cost manufacturing [19], [74]. Fabrication reliability of large-scale integrated systems could be greatly enhanced by the fault mitigation techniques discussed in Section IV-B2. While many other aspects of feasibility were not considered here, we have attempted to address the most common issues faced by prior optical information processing systems and believe the potential benefits of appropriately implementing a neuromorphic paradigm in an integrated optical platform constitute a reasonably compelling motivation for further investigation of photonic spike processing.

V. Conclusion

We have proposed a simple integrated scheme for parallel photonic neural interconnects called broadcast-and-weight, which exhibits properties unique among neuromorphic processors. The broadcast-and-weight architecture draws together principles of fiber optic communication, techniques of computational neuroscience, and recent technical advances in photonic system manufacturing. A reconfigurable PNN was proposed to grant networking functionality to a recently developed excitable laser processor, which behaves dynamically like a spiking neuron model. The PNN is a circuit method: it can be implemented with existing standard devices but could generalize to incorporate more advanced technologies, or even electronic dynamical units. By combining spike processing with WDM, a BL network exhibits a spatial flexibility that enables scalable spectrum reuse with great potential for organizational variety. An architecture of interfaced BLs appears to address many of the challenges encountered in prior proposals for scalable and feasible optical information processing, due in large part to particular correspondences between physical processes in optoelectronics and behavioral functions in the spiking model.

The present work solicits several possible directions for further research. To determine power use requirements, the sources and effects of noise on excitable laser neurons must be characterized. Adaptive control—both external and unsupervised—of many filter degrees of freedom represents a significant engineering challenge. Development of an untapped regime of high-speed, high-complexity processing would also call for development of applications and corresponding distributed algorithms, which could incorporate ideas from computational neuroscience and complex systems theory. The demands of these algorithms could further concretize design rules for multi-BL organization. The proposed architectural principles reveal an expansive scope of further challenges, yet they may represent a small step towards an unfamiliar and compelling model of photonic spike processing.

References

[1] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, no. 1–3, pp. 239–255, 2010.

[2] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Netw.*, vol. 10, no. 9, pp. 1659–1671, 1997.

[3] R. Sarpeshkar, "Analog versus digital: Extrapolating from electronics to neurobiology," *Neural Comput.*, vol. 10, pp. 1601–1638, 1998.

[4] S. Thorpe, A. Delorme, R. Van Rullen *et al.*, "Spike-based strategies for rapid processing," *Neural Netw.*, vol. 14, no. 6–7, pp. 715–725, 2001.

[5] 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, pp. 2531–2560, 2002.

[6] C. Savin, P. Joshi, and J. Triesch, "Independent component analysis in spiking neurons," *PLoS Comput. Biol.*, vol. 6, no. 4, p. e1000757, Apr. 2010.

[7] E. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. Cambridge, MA, USA: MIT Press, 2006.

[8] B. Szatmáry and E. M. Izhikevich, "Spike-timing theory of working memory," *PLoS Comput. Biol.*, vol. 6, no. 8, p. e1000879, 2010.

[9] J. Hasler and H. B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Frontiers Neurosci.*, vol. 7, no. 118, pp. 1–29, 2013.

[10] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Hfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen, "Neuromorphic silicon neuron circuits," *Frontiers Neurosci.*, vol. 5, no. 73, pp. 1–23, 2011.

[11] J. Seo, B. Brezzo, Y. Liu, B. Parker, S. Esser, R. Montoye, B. Rajendran, J. Tierno, L. Chang, D. Modha, and D. Friedman, "A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2011, pp. 1–4.

[12] M. D. Pickett, G. Medeiros-Ribeiro, and R. S. Williams, "A scalable neuristor built with Mott memristors," *Nature Mater.*, vol. 12, no. 2, pp. 114–117, 2013.

[13] K. Boahen, "Neurogrid: Emulating a million neurons in the cortex," presented at the IEEE EMBS Annu. Int. Conf., NY, USA, 2006.

[14] N. Imam, F. Akopyan, J. Arthur, P. Merolla, R. Manohar, and D. Modha, "A digital neurosynaptic core using event-driven QDI circuits," in *Proc. 18th IEEE Int. Symp. Asynchronous Circuits Syst.*, May 2012, pp. 25–32.

[15] S. Furber, D. Lester, L. Plana, J. Garside, E. Painkras, S. Temple, and A. Brown, "Overview of the SpiNNaker system architecture," *IEEE Trans. Comput.*, vol. 62, no. 12, pp. 2454–2467, Dec. 2013.

[16] J. Schemmel, J. Fieres, and K. Meier, "Wafer-scale integration of analog neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 431–438.

[17] B. Jalali and S. Fathpour, "Silicon photonics," *J. Lightw. Technol.*, vol. 24, no. 12, pp. 4600–4615, Dec. 2006.

[18] H. Park, A. W. Fang, D. Liang, Y.-H. Kuo, H.-H. Chang, B. R. Koch, H.-W. Chen, M. N. Sysak, R. Jones, and J. E. Bowers, "Photonic integration on the hybrid silicon evanescent device platform," *Adv. Opt. Technol.*, vol. 2008, article 682978, 2008.

[19] D. Liang, G. Roelkens, R. Baets, and J. E. Bowers, "Hybrid integrated platforms for silicon photonics," *Materials*, vol. 3, no. 3, pp. 1782–1802, 2010.

[20] H.-W. Chen, A. Fang, J. Peters, Z. Wang, J. Bovington, D. Liang, and J. Bowers, "Integrated microwave photonic filter on a hybrid silicon platform," *IEEE Trans. Microw. Theory Technol.*, vol. 58, no. 11, pp. 3213–3219, Nov. 2010.

[21] D. Rosenbluth, K. Kravtsov, M. P. Fok, and P. R. Prucnal, "A high performance photonic pulse processing device," *Opt. Exp.*, vol. 17, no. 25, pp. 22 767–22 772, 2009.

[22] M. P. Fok, H. Deming, M. Nahmias, N. Rafidi, D. Rosenbluth, A. Tait, Y. Tian, and P. R. Prucnal, "Signal feature recognition based on lightwave neuromorphic signal processing," *Opt. Lett.*, vol. 36, no. 1, pp. 19–21, Jan. 2011.

[23] M. P. Fok, Y. Tian, D. Rosenbluth, and P. R. Prucnal, "Asynchronous spiking photonic neuron for lightwave neuromorphic signal processing," *Opt. Lett.*, vol. 37, no. 16, pp. 3309–3311, Aug. 2012.

[24] Y. Tian, M. P. Fok, D. Rosenbluth, and P. Prucnal, "Asynchronous spiking neuron based on four-wave mixing and cross absorption modulation," presented at the Opt. Fiber Commun. Conf., Los Angeles, CA, USA, 2012, Paper OTh3H.1.

[25] A. N. Tait, M. A. Nahmias, Y. Tian, B. J. Shastri, and P. R. Prucnal, "Photonic neuromorphic signal processing and computing," in *Nanophotonic Information Physics*. Berlin, Germany: Springer, 2014, pp. 183–222.

[26] 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*, vol. 84, no. 3, p. 036209, 2011.

[27] 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. Exp.*, vol. 20, no. 18, pp. 20292–20308, Aug. 2012.

[28] B. Shastri, M. Nahmias, A. Tait, Y. Tian, M. Fok, M. Chang, B. Wu, and P. Prucnal, "Exploring excitability in graphene for spike processing networks," in *Proc. 13th Int. Conf. Numerical Simul. Optoelectron. Devices.*,Aug. 2013, pp. 83–84.

[29] A. Hurtado, K. Schires, I. Henning, and M. Adams, "Investigation of vertical cavity surface emitting laser dynamics for neuromorphic photonic systems," *Appl. Phys. Lett.*, vol. 100, no. 10, pp. 103 703–103 703, 2012.

[30] B. J. Shastri, M. A. Nahmias, A. N. Tait, Y. Tian, B. Wu, and P. R. Prucnal, "Graphene excitable laser for photonic spike processing," presented at the IEEE Photon. Conf., Seattle, WA, USA, Sep. 2013, pp. 1–2, Paper PD.4.

[31] M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," *J. Sel. Topics Quantum Electron.*, vol. 19, no. 1800212, pp. 1–12, 2013.

[32] M. A. Nahmias, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "An evanescent hybrid silicon laser neuron," presented at the IEEE Photon. Conf.. Seattle, WA, USA, Sep. 2013, pp. 93–94, Paper ME3.4.

[33] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast-and-weight interconnects for integrated distributed processing systems," in *Proc. IEEE Opt. Interconnects Conf.*, May 2014.

[34] J. Psota, J. Miller, G. Kurian, H. Hoffman, N. Beckmann, J. Eastep, and A. Agarwal, "ATAC: Improving performance and programmability with on-chip optical networks," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2010, pp. 3325–3328.

[35] S. Le Beux, J. Trajkovic, I. O'Connor, G. Nicolescu, G. Bois, and P. Paulin, "Optical ring network-on-chip (ORNoC): Architecture and design methodology," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, 2011, pp. 1–6.

[36] S. Rakheja and V. Kumar, "Comparison of electrical, optical and plasmonic on-chip interconnects based on delay and energy considerations," in *Proc. 13th Int. Symp. Quality Electron. Des.*, 2012, pp. 732–739.

[37] Q. Xu and M. Lipson, "All-optical logic based on silicon micro-ring resonators," *Opt. Exp.*, vol. 15, no. 3, pp. 924–929, Feb. 2007.

[38] D. Sridharan and E. Waks, "All-optical switch using quantum-dot saturable absorbers in a DBR microcavity," *IEEE J. Quantum Electron.*, vol. 47, no. 1, pp. 31–39, Jan. 2011.

[39] R. W. Keyes, "Optical logic-in the light of computer technology," *Optica Acta: Int. J. Opt.*, vol. 32, no. 5, pp. 525–535, 1985.

[40] D. A. B. Miller, "Are optical transistors the logical next step?" *Nature Photon.*, vol. 4, no. 1, pp. 3–5, Jan. 2010.

[41] A. Tait, B. Shastri, M. Fok, M. Nahmias, and P. Prucnal, "The DREAM: An integrated photonic thresholder," *J. Lightw. Technol.*, vol. 31, no. 8, pp. 1263–1272, Apr. 2013.

[42] M. Aono, M. Naruse, S.-J. Kim, M. Wakabayashi, H. Hori, M. Ohtsu, and M. Hara, "Amoeba-inspired nanoarchitectonic computing: Solving intractable computational problems using nanoscale photoexcitation transfer dynamics," *Langmuir*, vol. 29, no. 24, pp. 7557–7564, 2013.

[43] L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer, "Photonic information processing beyond turing: An optoelectronic implementation of reservoir computing," *Opt. Exp.*, vol. 20, no. 3, pp. 3241–3249, Jan. 2012.

[44] Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, "Optoelectronic reservoir computing," *Sci. Rep.*, vol. 2, no. 287, pp. 1–6, Feb. 2012.

[45] M. Naruse, N. Tate, M. Aono, and M. Ohtsu, "Information physics fundamentals of nanophotonics," *Rep. Progress Phys.*, vol. 76, no. 5, p. 056401, 2013.

[46] S. Song, K. D. Miller, and L. F. Abbott, 'Competitive Hebbian learning through spike-timing-dependent synaptic plasticity," *Nature: Neurosci.*, vol. 3, no. 9, pp. 919–926, 2000.

[47] M. P. Fok, Y. Tian, D. Rosenbluth, and P. R. Prucnal, "Pulse lead/lag timing detection for adaptive feedback and control based on optical spike-timing-dependent plasticity," *Opt. Lett.*, vol. 38, no. 4, pp. 419–421, Feb. 2013.

[48] E. C. Mos, J. J. H. B. Schleipen, H. de Waardt, and D. G. D. Khoe, "Loop mirror laser neural network with a fast liquid-crystal display," *Appl. Opt.*, vol. 38, no. 20, pp. 4359–4368, Jul.1999.

[49] S. L. Yeh, R. C. Lo, and C. Y. Shi, "Optical implementation of the Hopfield neural network with matrix gratings," *Appl. Opt.*, vol. 43, no. 4, pp. 858–865, Feb. 2004.

[50] P. Asthana, G. P. Nordin, J. Armand R. Tanguay, and B. K. Jenkins, "Analysis of weighted fan-out/fan-in volume holographic optical interconnections," *Appl. Opt.*, vol. 32, no. 8, pp. 1441–1469, Mar. 1993.

[51] J. Shamir, H. J. Caulfield, and R. B. Johnson, "Massive holographic interconnection networks and their limitations," *Appl. Opt.*, vol. 28, no. 2, pp. 311–324, Jan.1989.

[52] J. W. Goodman, "Fan-in and fan-out with optical interconnections," *Opt. Acta: Int. J. Opt.*, vol. 32, no. 12, pp. 1489–1496, 1985.

[53] M. Hill, E. E. E. Frietman, H. de Waardt, G.-D. Khoe, and H. Dorren, "All fiber-optic neural network using coupled SOA based ring lasers," *IEEE Trans. Neural Netw.*, vol. 13, no. 6, pp. 1504–1513, Nov. 2002.

[54] K. Preston, N. Sherwood-Droz, J. S. Levy, and M. Lipson, "Performance guidelines for WDM interconnects based on silicon microring resonators," presented at the CLEO: Sci. Innovation., Baltimore, MD, USA, 2011, Paper CThP4.

[55] R. Ramaswami, "Multiwavelength lightwave networks for computer communication," *IEEE Commun. Mag.*, vol. 31, no. 2, pp. 78–88, Feb.1993.

[56] T. Wood and N. K. Shankaranarayanan, "Operation of a passive optical network with subcarrier multiplexing in the presence of optical beat interference," *J. Lightw. Technol.*, vol. 11, no. 10, pp. 1632–1640, Oct.1993.

[57] Q. Xu and R. Soref, "Reconfigurable optical directed-logic circuits using microresonator-based optical switches," *Opt. Exp.*, vol. 19, no. 6, pp. 5244–5259, Mar. 2011.

[58] J. Chang, Y. Deng, M. P. Fok, J. Meister, and P. R. Prucnal, "Photonic microwave finite impulse response filter using a spectrally sliced supercontinuum source," *Appl. Opt.*, vol. 51, no. 19, pp. 4265–4268, Jul. 2012.

[59] A. Fang, M. Sysak, B. Koch, R. Jones, E. Lively, Y. Hao Kuo, D. Liang, O. Raday, and J. Bowers, "Single-wavelength silicon evanescent lasers," *IEEE J. Sel. Topics Quantum Electron.*, vol. 15, no. 3, pp. 535–544, Nov. 2009.

[60] X. Zhang, J. Wei, and C. Qiao, "Constrained multicast routing in WDM networks with sparse light splitting," *J. Lightw. Technol.*, vol. 18, no. 12, pp. 1917–1927, Dec. 2000.

[61] G. Duan, C. Jany, A. Le Liepvre, M. Lamponi, A. Accard, F. Poingt, D. Make, F. Lelarge, S. Messaoudene, D. Bordel, J. Fedeli, S. Keyvaninia, G. Roelkens, D. Van Thourhout, D. Thomson, F. Gardes, and G. Reed, "Integrated hybrid III-V/Si laser and transmitter," in *Proc. Int. Conf. Indium Phosphide Related Mater.*, Aug. 2012, pp. 16–19.

[62] K. K. Lee, D. R. Lim, L. C. Kimerling, J. Shin, and F. Cerrina, "Fabrication of ultralow-loss si/sio2 waveguides by roughness reduction," *Opt. Lett.*, vol. 26, no. 23, pp. 1888–1890, Dec. 2001.

[63] P. Christie and D. Stroobandt, "The interpretation and application of Rent's rule," *IEEE Trans. Very Large Scale Integr.*, vol. 8, no. 6, pp. 639–648, Dec. 2000.

[64] D. Miller, "Device requirements for optical interconnects to silicon chips," *Proc. IEEE*, vol. 97, no. 7, pp. 1166–1185, Jul. 2009.

[65] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[66] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[67] M. Shanahan, "Dynamical complexity in small-world networks of spiking neurons," *Phys. Rev. E*, vol. 78, p. 041924, Oct. 2008.

[68] D. S. Bassett and E. Bullmore, "Small-world brain networks," *Neuroscientist*, vol. 12, no. 6, pp. 512–523, 2006.

[69] D. Meunier, R. Lambiotte, and E. T. Bullmore, "Modular and hierarchically modular organization of brain networks," *Frontiers Neurosci.*, vol. 4, no. 200, pp. 1–11, 2010.

[70] J. Hardy and J. Shamir, "Optics inspired logic architecture," *Opt. Exp.*, vol. 15, no. 1, pp. 150–165, Jan. 2007.

[71] E. M. Izhikevich, "Polychronization: Computation with spikes," *Neural Comput.*, vol. 18, pp. 245–282, 2006.

[72] R. Reischuk, "Can large fan-in circuits perform reliable computations in the presence of noise?" in *Computing and Combinatorics*. New York, NY, USA: Springer, 1997, pp. 72–81.

[73] G. S. Snider, "Self-organized computation with unreliable, memristive nanodevices," *Nanotechnology*, vol. 18, no. 36, p. 365202, 2007.

[74] R. Soref, "The past, present, and future of silicon photonics," *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, no. 6, pp. 1678–1687, Nov./Dec. 2006.

Alexander N. Tait (S'11) received the B.Sci.Eng. (Hons.) in electrical engineering from Princeton University, Princeton, NJ, USA, in 2012, where he is currently working toward the Ph.D. degree in electrical engineering in the Lightwave Communications Group, Department of Electrical Engineering.

He was a Research Intern for the summers of 2008–2010 at the Laboratory for Laser Energetics, University of Rochester, Rochester, NY, USA, and an Undergraduate Researcher for the summers of 2011–2012 at the MIRTHE Center, Princeton University, Princeton, NJ, USA. His research interests include photonic devices for nonlinear signal processing, integrated systems, neuromorphic engineering, and hybrid analog–digital signal processing and computing.

Mr. Tait is a Student Member of the IEEE Photonics Society and the Optical Society of America. He received the National Science Foundation Graduate Research Fellowship. He received the Optical Engineering Award of Excellence from the Department of Electrical Engineering, Princeton University. He has coauthored six journal papers and one Springer book chapter.

Mitchell A. Nahmias (S'11) Graduated (Hons.) from Princeton University, Princeton, NJ, USA, with the B.S. degree in electrical engineering and a certificate in engineering physics. He is currently working toward the Ph.D. degree in electrical engineering under Prof. P. Prucnal to continue his undergraduate work on his excitable, photonic neuron.

Mr. Nahmias received the John Ogden Bigelow Jr. Prize in Electrical Engineering and Cowinner of the "Best Engineering Physics Independent Work Award" for his senior thesis. He received the National Science Foundation Graduate Research Fellowship.

Bhavin J. Shastri (S'03–M'11) received the B.Eng. (Hons. with distinction), M.Eng., and Ph.D. degrees in electrical engineering from McGill University, Montreal, QC, Canada, in 2005, 2007, and 2011, respectively.

He is currently a Postdoctoral Research Fellow at the Lightwave Communications Laboratory, Princeton University, Princeton, NJ, USA. His research interests include ultrafast cognitive computing—neuromorphic engineering with photonic neurons, high-speed burst-mode clock and data recovery circuits, optoelectronic-VLSI systems, optical access networks, machine learning, and computer vision.

Dr. Shastri has received the following research awards: 2012 D. W. Ambridge Prize for the top graduating Ph.D. student, nomination for the 2012 Canadian Governor General Gold Medal, IEEE Photonics Society 2011 Graduate Student Fellowship, 2011 Postdoctoral Fellowship from National Sciences and Engineering Research Council of Canada (NSERC), 2011 SPIE Scholarship in Optics and Photonics, a Lorne Trottier Engineering Graduate Fellow, and a 2008 Alexander Graham Bell Canada Graduate Scholarship from NSERC. He received the Best Student Paper Award at the 2010 IEEE Midwest Symposium on Circuits and Systems, the corecipient of the Silver Leaf Certificate at the 2008 IEEE Microsystems and Nanoelectronics Conference, the 2004 IEEE Computer Society Lance Stafford Larson Outstanding Student Award, and the 2003 IEEE Canada Life Member Award. He was the President/Cofounder of the McGill OSA Student Chapter.

Paul R. Prucnal (S'75–M'79–SM'90–F'92) received the A.B. degree from Bowdoin College (*summa cum laude*), with Highest Honors in math and physics, where he was elected to Phi Beta Kappa. He received the M.S., M.Phil., and Ph.D. degrees from Columbia University, where he was elected to the Sigma Xi Honor Society. He is currently a Professor of Electrical Engineering, Princeton University, Princeton, NJ, USA, where he has also served as the Founding Director of the Center for Photonics and Optoelectronic Materials. He has held visiting faculty positions at the University of Tokyo and University of Parma.

Prof. Prucnal was an Area Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS for optical networks, and was Technical Chair and General Chair of the IEEE Topical Meeting on Photonics in Switching in 1997 and 1999, respectively. He is a Fellow of IEEE with reference to his work on optical networks and photonic switching, a Fellow of the OSA, and a recipient of the Rudolf Kingslake Medal from the SPIE, cited for his seminal paper on photonic switching. In 2006, he received the Gold Medal from the Faculty of Physics, Mathematics and Optics from Comenius University in Slovakia, for his contributions to research in photonics. He has received Princeton Engineering Council Awards for Excellence in Teaching, the University Graduate Mentoring Award, and the Walter Curtis Johnson Prize for Teaching Excellence in Electrical Engineering, as well as the Distinguished Teacher Award from Princetons School of Engineering and Applied Science. He is Editor of the book, Optical Code Division Multiple Access: Fundamentals and Applications.Ó