Review article **Open Access**

Thomas Ferreira de Lima*, Bhavin J. Shastri, Alexander N. Tait, Mitchell A. Nahmias and Paul R. Prucnal

Progress in neuromorphic photonics

DOI 10.1515/nanoph-2016-0139
Received August 5, 2016; revised October 31, 2016; accepted November 21, 2016

Abstract: As society's appetite for information continues to grow, so does our need to process this information with increasing speed and versatility. Many believe that the one-size-fits-all solution of digital electronics is becoming a limiting factor in certain areas such as data links, cognitive radio, and ultrafast control. Analog photonic devices have found relatively simple signal processing niches where electronics can no longer provide sufficient speed and reconfigurability. Recently, the landscape for commercially manufacturable photonic chips has been changing rapidly and now promises to achieve economies of scale previously enjoyed solely by microelectronics. By bridging the mathematical prowess of artificial neural networks to the underlying physics of optoelectronic devices, neuromorphic photonics could breach new domains of information processing demanding significant complexity, low cost, and unmatched speed. In this article, we review the progress in neuromorphic photonics, focusing on photonic integrated devices. The challenges and design rules for optoelectronic instantiation of artificial neurons are presented. The proposed photonic architecture revolves around the processing network node composed of two parts: a nonlinear element and a network interface. We then survey excitable lasers in the recent literature as candidates for the nonlinear node and microring-resonator weight banks as the network interface. Finally, we compare metrics between neuromorphic electronics and neuromorphic photonics and discuss potential applications.

Keywords: neuromorphic computing; photonic integrated circuits; ultrafast information processing; excitable semiconductor lasers.

***Corresponding author: Thomas Ferreira de Lima,** Lightwave Communications Research Laboratory, Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA, e-mail: tlima@princeton.edu
Bhavin J. Shastri, Alexander N. Tait, Mitchell A. Nahmias and Paul R. Prucnal: Lightwave Communications Research Laboratory, Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

1 Introduction to neuromorphic engineering

The success of digital electronics has created a data-hungry consumer society, which in turn reinvested in more capable, faster, and cheaper machines. For decades, the transistor count of CPUs doubled every 2 years, a trend that became known as Moore's law. Microprocessor clock rates also increased exponentially, but current leakage in nanometric nodes became prevalent, causing a halt to this growth at about 4 GHz [1]. At the same time, the past decade has seen the breakdown of Dennard scaling [2]; the power density of microelectronic chips no longer stays constant as they get denser, that is, smaller transistors do not consume less power. The recent shift to multicore scaling alleviated these constraints, but the breakdown of Dennard scaling has limited the number of cores than can simultaneously be powered on with a fixed power budget and heat extraction rate, giving rise to the "dark silicon" phenomenon [3]. Projections for the 8 nm node indicate that more than 50% of the chip will be "dark" [3]. Fundamentally, these issues can be traced to two primary physical bottlenecks: the bandwidth limitations of metal interconnects and the energy consumption and subsequently heat generation of digital switching [4]. In summary, operating speed and power efficiency of CPUs have reached physical barriers that cannot be addressed through Dennard scaling. Consequently, this has opened up new opportunities in unconventional information processing architectures, which include an array of different processing modalities [5].

The computational efficiency, measured in joules per MAC (multiply and accumulate operation, as revisited in Section 5), have been scaling similarly (Koomey's law), but it has slowed down significantly in the last few years; it has only improved by a factor of about two over the last 14 years, and it is now reaching an asymptotic power efficiency wall of about 100 pJ/MAC.

Respecting power budgets is now a top priority for digital processors. Data centers, Wi-Fi routers, and Internet traffic represent a tremendous electric energy consumption. Current trends indicate a shift of electricity usage from consumer device use to network and data centers [6, 7]. In the worst-case scenario, at the rate at which societal consumption and production of data is growing, it is predicted that fixed-access networks (Wi-Fi and LAN) and data centers will consume up to 33% of the world's energy use [6].

To counter that trend, power-aware large-scale integration techniques in photonics are just emerging, being pushed forward by data communication applications and a market need for increased information flow between processors, on both macro and micro scales [7, 8]. This has led to an explosion in photonic integrated circuits (PIC), which are already finding their way into fast Ethernet switches for servers and supercomputers and will likely emerge in more traditional processor architectures as electronic interconnects fail to keep up with data volume. The average energy efficiency of the world's fastest supercomputers lies in the order of 1 nJ/FLOP [Gre], where FLOP stands for floating-point operation, a standard computing unit. In green data centers and high-performance computers, there is an urgent need for unconventional, special-purpose coprocessors with efficiencies beyond 1 nJ/FLOP, with a caveat: these coprocessors must operate at the same throughput handled by the high-speed digital and analog circuits it interfaces with, so they do not become a bottleneck.

This efficiency level is not fundamentally impossible. In fact, the human brain is estimated to being able to compute an amazing 10^{20} MAC/s using only 20 W of power [9] [MAC operation; cf. Section 5, similar to FLOP but more appropriate for digital signal processors (DSP)]. It does this with 10^{11} neurons with spike firing rate of ~1 Hz but with a large number of interconnects per neuron (10^4), highlighting the importance of distributed processing (see Section 2.1). The calculated computational efficiency for the brain is therefore nine orders of magnitude beyond that of current supercomputers (<aJ/MAC). "Neuromorphic computing" offers hope to building large-scale "bioinspired" hardware for specialized processing while attempting computational efficiencies past the von Neumann efficiency wall toward those of a human brain.

1.1 Neuromorphic microelectronics

Various technologies have demonstrated large-scale spiking neural networks (SNNs) in electronics, including, notably, Neurogrid as part of Stanford University's Brains in Silicon program [10], IBM's TrueNorth as part of the Defense Advanced Research Projects Agency's (DARPA) SyNAPSE program [11], HICANN as part of the University of Heidelberg's FACETS/BrainScaleS project [12], and University of Manchester's neuromorphic chip as part of the SpiNNaker project [13]; the latter two are under the flagship of the European Commission's Human Brain Project [14].

Whereas von Neumann processors rely on a single point-to-point link between memory and CPU, a neuromorphic processor typically requires a large number of interconnects (i.e. ~100 s of many-to-one fan-in per processor) [9]. This requires a significant amount of multicasting, resulting in a distributed communication burden. This, in turn, introduces fundamental performance challenges that result from capacitive loads and radiative physics in electronic links in addition to the typical bandwidth-distance-energy limits of point-to-point connections [15]. Realistically scalable systems are ultimately forced to adopt a combination of crossbar time-division multiplexing (TDM) and/or packet switching (e.g. [11]). Address-event representation (AER) introduces the overhead of representing spike as digital codes instead of physical pulses. This abstraction at the architectural level allows virtual interconnectivity to exceed wire density by a factor related to the sacrificed bandwidth, which can be orders of magnitude [16]. SNNs based on AER are thus effective at targeting biological timescales and the associated application space: real-time applications (object recognition) in the kHz regime [11, 13] and accelerated simulation in the low MHz regime [12]. However, neuromorphic processing for high-bandwidth applications that require GHz operation per neuron (such as sensing and manipulating the radiospectrum and for hypersonic aircraft control) must take a fundamentally different approach to interconnection.

1.2 Why neuromorphic photonics?

Photonics has revolutionized information transmission (communication and interconnects), whereas electronics, in parallel, has dominated information transformation (computation). This leads naturally to the following question: how can we unify the advantages of the two as effectively as possible? [17]. CMOS gates only draw energy from the rail when and where called upon; however, the energy required to driving an interconnect from one gate to the next dominates CMOS circuit energy use. Relaying a signal from gate to gate, especially using a clocked

scheme, induces penalties in latency and bandwidth compared to an optical waveguide passively carrying multiplexed signals.

This suggests that starting up a new architecture from a photonic interconnection fabric supporting nonlinear optoelectronic devices can be uniquely advantageous in terms of energy efficiency, bandwidth, and latency, sidestepping many of the fundamental trade-offs in digital and analog electronics. It may be one of the few practical ways to achieve ultrafast, complex on-chip processing without consuming impractical amounts of power [18].

Complex photonic systems have been largely unexplored due to the absence of a robust photonic integration industry. Recently, however, the landscape for manufacturable photonic chips has been changing rapidly and now promises to achieve economies of scale previously enjoyed solely by microelectronics. In particular, a new photonic manufacturing hybrid platform that combines in the same chip both active elements (e.g. lasers and detectors) and passive elements (e.g. waveguides, resonators, and modulators) is emerging [19]. A neuromorphic photonic approach based on this platform could potentially operate 6–8 orders of magnitude faster than neuromorphic electronics when accounting for the bandwidth reduction of virtualizing interconnects [20] (cf. Figure 1).

1.3 Emergence of neuromorphic photonics

The key criteria for nonlinear elements to enable a scalable computing platform include [17] thresholding, fan-in, and cascadability. Past approaches to optical computing have met challenges realizing these criteria; so far, no optical logic device satisfying all of them has been proposed. More recent investigations, introduced in the following sections, have concluded that a photonic neuromorphic processor could satisfy them by implementing a model of a neuron as opposed to the model of a logic gate.

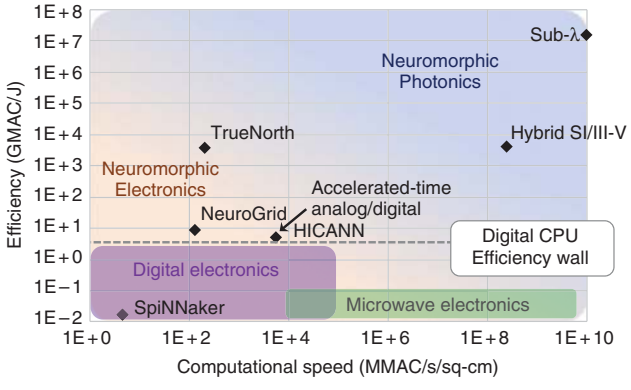

Figure 1: Speed and efficiency metrics that are accessible by various neuromorphic hardware platforms.
On the top right, the two photonic neuron platforms studied in Ref. [20]: hybrid III-V/Si stands for III-V/silicon hybrid platform SNN PIC. Sub-λ stands for subwavelength photonics. The other points refer to the recent electronic neuromorphic hardware, as discussed in Section 5. The regions highlighted in the graph are approximate based on qualitative trade-offs of each technology.

Early work in neuromorphic photonics involved fiber-based spiking approaches for learning, pattern recognition, and feedback [21–23]. Spiking behavior resulted from a combination of semiconductor optical amplifiers (SOA) together with a highly nonlinear fiber thresholder, but they were neither excitable nor asynchronous and therefore not suitable for scalable, distributed processing in networks.

"Neuromorphism" implies a strict isomorphism between artificial neural networks and optoelectronic devices (Section 2). There are two research challenges necessary to establish this isomorphism: the nonlinearity (equivalent to thresholding) in individual neurons, as discussed in Section 3, and the synaptic interconnection (related to fan-in and cascadability) between different neurons, as discussed in Section 4. Once the isomorphism is established and large networks are fabricated, we anticipate that the computational neuroscience and software engineering will have a new optimized processor for which they can adapt their methods and algorithms (cf. Section 6).

Recent investigations have concluded that a photonic subcircuit called the processing network node (PNN) could satisfy them by implementing a model of a neuron as opposed to the model of a logic gate.

Photonic unconventional computing primitives such as the PNN (Section 3) address the traditional problem of noise accumulation by interleaving physical representations of information. Representational interleaving, in which a signal is repeatedly transformed between coding schemes (digital-analog) or physical variables (electronic-optical), can grant many advantages to computation and noise properties. From an engineering standpoint, the logical function of a nonlinear neuron can be thought of as increasing signal-to-noise ratio (SNR) that tends to degrade in linear systems, whether that means a continuous nonlinear transfer function suppressing analog noise or spiking dynamics curtailing pulse attenuation and spreading. As a result, we neglect purely linear PNNs as they do not offer mechanisms to maintain signal fidelity in a large network in the presence of noise.

The optical channel alone is highly expressive and correspondingly very sensitive to phase and frequency

noise. For example, the networking architecture proposed in Section 4 relies on wavelength-division multiplexing (WDM) for interconnecting many points in a photonic substrate together. Any proposal for networking computational must address the issue of practical cascadability: transferring information and energy in the optical domain from one neuron to many others and exciting them with the same strength without being sensitive to noise. This is notably achieved, for example, by encoding information in energy pulses that can trigger stereotypical excitation in other neurons regardless of their analog amplitude.

In this article, we review the progress in neuromorphic photonics research, focusing especially on integrated photonic devices. An elegant parallel between neural networks and optoelectronic devices such as excitable lasers can be established and exploited for processing. Section 2 introduces the concept of a "photonic neuron" followed by a discussion on its feasibility. Then, Section 3 presents a review on recent research on optical devices that could be used as a primitive node in photonic neural networks. Section 4 presents a networking architecture that efficiently channelizes the spectrum of an integrated waveguide. Finally, Section 5 provides a quantitative analysis of neuromorphic photonics in the context of electronic approaches.

2 Photonic neuron

2.1 What is an artificial neuron?

Neuroscientists research artificial neural networks as an attempt to mimic the "natural processing" capabilities of the brain. These networks of simple nonlinear nodes can be taught (rather than programmed) and reconfigured to best execute a desired task; this is called "learning". Today, neural nets offer state-of-the-art algorithms for machine intelligence such as speech recognition, natural language processing, and machine vision [24].

Three elements constitute a neural network: a set of nonlinear nodes (neurons), configurable interconnection (network), and information representation (coding scheme). An elementary illustration of a neuron is shown in Figure 2. The network consists of a weighted directed graph, in which connections are called synapses. The input of a neuron is a linear combination (or weighted addition) of the outputs of the neurons connected to it. Then, the particular neuron integrates the combined signal and produces a nonlinear response, as represented by an "activation function", usually monotonic and bounded.

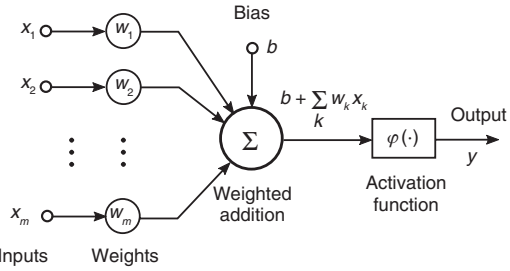

Figure 2: Nonlinear model of a neuron.
Note the three parts: (i) a set of "synapses" or "connecting links"; (ii) an "adder" or "linear combiner", performing weighted addition; and (iii) a nonlinear "activation function". From Ref. [18].

Three generations of neural networks were historically studied in computational neuroscience [25]. The first was based on the McCulloch-Pitts neural model, which consists of a linear combiner followed by a step-like activation function (binary output). These neural networks are Boolean-complete, that is, they have the ability of simulating any Boolean circuit and are said to be universal for digital computations. The second generation implemented analog outputs, with a continuous activation function instead of a hard thresholder. Neural networks of the second generation are universal for analog computations in the sense that they can uniformly approximate arbitrarily well any continuous function with a compact domain [25]. When augmented with the notion of "time", recurrent connections can be created and exploited to create attractor states [26] and associative memory [27] in the network.

Physiological neurons communicate with each other using pulses called action potentials or spikes. In traditional neural network models, an analog variable is used to represent the firing rate of these spikes. This coding scheme called "rate coding" was believed to be a major, if not the only, coding scheme used in biology. Surprisingly, there are some fast analog computations in the visual cortex that cannot possibly be explained by rate coding. For example, neuroscientists demonstrated in the 1990s that a single cortical area in macaque monkeys is capable of analyzing and classifying visual patterns in just 30 ms in spite of the fact that these neurons' firing rates are usually below 100 Hz (i.e. less than 3 spikes in 30 ms) [25, 28, 29], which directly challenges the assumptions of rate coding. In parallel, more evidence was found that biological neurons use the precise timing of these spikes to encode information, which led to the investigation of a third generation of neural networks based on a "spiking neuron".

The simplicity of the models of the previous generations precluded the investigation of the possibilities of using "time" as resource for computation and communication. If the "timing" of individual spikes itself carry analog information ("temporal coding"), then the energy necessary to create such spike is optimally employed to express information. Furthermore, Maass showed that this third generation is a generalization of the first two and, for several concrete examples, can emulate real-valued neural network models while being more robust to noise [25].

For example, one of the simplest models of a spiking neuron is called "leaky integrate-and-fire" (LIF), as described in Eq. (1). It represents a simplified circuit model of the membrane potential of a biological spiking neuron.

$$C_{\mathrm{m}}\frac{\mathrm{d}V_{\mathrm{m}}(t)}{\mathrm{d}t}=-\frac{1}{R_{\mathrm{m}}}(V_{\mathrm{m}}(t)-V_{\mathrm{L}})+I_{\mathrm{app}}(t); \tag{1}$$

if $V_{\mathrm{m}}(t)>V_{\mathrm{thresh}}$, then release a spike and set $V_{\mathrm{m}}(t)\rightarrow V_{\mathrm{reset}}$, where $V_{\mathrm{m}}(t)$ is the membrane voltage, R_{m} is the membrane resistance, V_{L} is the equilibrium potential, and I_{app} is the applied current (input). More biorealistic models, such as the Hodgkin-Huxley model, involve several ordinary differential equations and nonlinear functions.

However, simply simulating neural networks on a conventional computer, be it of any generation, is costly because of the fundamentally serial nature of CPU architectures. Biorealistic SNNs present a particular challenge because of the need for fine-grained time discretization [30]. Engineers circumvent this challenge by employing an event-driven simulation model that resolves this issue by storing the time and shape of the events expanded in a suitable basis in a simulation queue. Although simplified models do not faithfully reproduce key properties of cortical spiking neurons, it allows for large-scale simulations of SNNs, from which key networking properties can be extracted. These costs defeat the purpose of using spiking neurons for engineering applications.

Alternatively, one can build an unconventional, distributed network of nonlinear nodes that directly use the physics of nonlinear devices or excitable dynamical systems, significantly dropping energetic cost per bit.

Here, we will discuss recent advances in neuromorphic photonic hardware and the constraints to which particular implementations must subject, including accuracy, noise, cascadability, and thresholding. A successful architecture must tolerate eventual inaccuracies and noise, indefinite propagation of signals, and provide mechanisms to counteract noise accumulation as the signal traverses across the network.

2.2 Basic requirements for a photonic neuron

An artificial neuron described in Figure 2 must perform three basic mathematical operations: vector multiplication (weighting), spatial summation (addition), and a nonlinear transformation (activation function). Moreover, the inputs to be weighted in the first stage must be of the same nature of the output – in the case considered here, photons.

As the size of the network grows, additional mechanisms are required at the hardware level to ensure the integrity of the signals. The neuron must have a scalable number of inputs, referred to as "maximum fan-in" (N_{f}), which will determine the degree of connectivity of the network. Each neuron's output power must be strong enough to drive at least N_{f} others ("cascadability"). This concept is tied closely with that of "thresholding": the SNR at the output must be lower than at its input. Cascadability, thresholding, and fan-in are particularly challenging to optical systems due to quantum efficiency (photons have finite supply) and amplified spontaneous emission (ASE) noise, which degrades SNR.

2.3 Photonic neuron module: PNN

A networkable photonic device with optical I/O, provided that it is capable of emulating an artificial neuron, is named a PNN [31]. Formulations of a photonic PNN can be divided into two main categories: all-optical and optical-electrical-optical (O/E/O), respectively classified according to whether the information is always embedded in the optical domain or switches from optical to electrical and back. We note that the term "all-optical" is sometimes very loosely defined in engineering articles. Physicists reserve it for devices that rely on parametric nonlinear processes, such as four-wave mixing. Here, our definition includes devices that undergo nonparametric processes as well, such as semiconductor lasers with optical feedback, in which optical pulses directly perturb the carrier population, triggering quick energy exchanges with the cavity field that results in the release of another optical pulse.

WDM efficiently uses the spectral window of optical waveguides, maximizing the information throughput in a single waveguide. Therefore, it is highly desirable and crucial to design a PNN that is compatible with WDM. All-optical versions of a PNN must have some way to sum multiwavelength signals, and this requires a population of charge carriers. On the contrary, O/E/O versions could make use of photodetectors (PD) to provide a spatial sum of

WDM signals. The PD output could drive an E/O converter, involving a laser or a modulator, whose optical output is a nonlinear result of the electrical input. Instances of both techniques are presented in Section 3.

2.3.1 All-optical PNNs

Coherent injection models are characterized by input signals directly interacting with cavity modes, such that outputs are at the same wavelength as inputs (Figure 3A). Because coherent optical systems operate at a single wavelength λ, the signals lack distinguishability from one another in a WDM-encoded framework. As demonstrated in Ref. [32], the effective weight of coherently injected inputs is also strongly phase dependent. Global optical phase control presents a challenge in synchronized laser systems but also affords an extra degree of freedom to configure weight values.

Incoherent injection models inject light in a wavelength λ_j to selectively modulate an intracavity property that then triggers excitable output pulses in an output wavelength λ_i (Figure 3B). A number of approaches [33–36], including those based on optical pumping, fall under this category. Although distinct, the output wavelength often has a stringent relationship with the input wavelength. For example, excitable micropillar lasers [35, 37] are carefully designed to support one input mode with a node coincident with an anti-node of the lasing mode. In cases where the input is also used as a pump [38–40], the input wavelength must be shorter than that of the output to achieve carrier population inversion.

WDM networking introduces wavelength constraints that conflict with the ones inherent to optical injection.

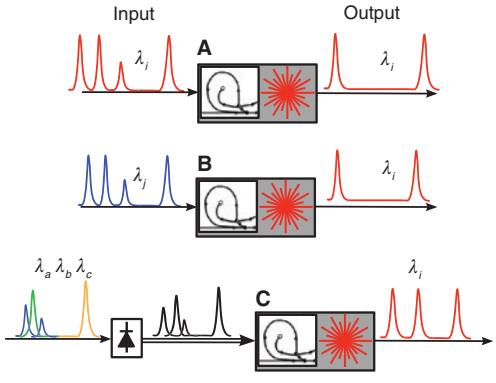

Figure 3: General classification of semiconductor excitable lasers based on (A) coherent optical injection electrical injection, (B) noncoherent optical injection, and (C) full electrical injection. Each of these lasers can be pumped either electrically or optically.

One approach for networking optically injected devices is to attempt to separate these wavelength constraints. In an early work on neuromorphic photonics in fiber, this was accomplished with charge-carrier-mediated cross-gain modulation (XGM) in an SOA [21–23].

2.3.2 O/E/O PNNs

In this kind of PNN, the O/E subcircuit is responsible for the weighted addition functionality, whereas the E/O is responsible for the nonlinearity (Figure 3C). Each subcircuit can therefore be analyzed independently. The analysis of an O/E WDM weighted addition circuit is referred to Section 4.

The E/O subcircuit of the PNN must take an electronic input representing the complementary weighted sum of optical inputs, perform some dynamical or nonlinear process, and generate a clean optical output on a single wavelength. Figure 4 classifies the six different ways in which nonlinearities can be implemented in an E/O circuit. The type of nonlinearity, corresponding to different neural models, is separated into "dynamical systems" and "continuous nonlinearities", both of which have a single input u and output y. A continuous nonlinearity is described by a differential equation $\dot{y} = f(y, u)$. This includes continuous-time recurrent neural networks (CTRNNs) such as Hopfield networks. The derivative of y introduces a sense of time, which is required to consider recurrent networking, although it does not exclude feedforward models where time plays no role, such as perceptron models. A dynamical system has an internal state $\vec{x}$ and is described by $\dot{\vec{x}} = g(\vec{x}, u); \dot{y} = h(\vec{x}, y, u)$, where the second differential equation represents the mapping between the internal state $\vec{x}$ and the output y. There are a wide variety of spiking models based on excitability, threshold behavior, and relaxation oscillations, covered, for example, in Ref. [43].

Physical implementations of these nonlinearities can arise from devices falling into roughly three categories: pure electronics, electro-optic physics in modulators, and active laser behavior (Figure 4). Figure 4A illustrates spiking lasers, which are detailed in Section 3 and offer perhaps the most promise in terms of garnering the full advantage of recent theoretical results on spike processing efficiency and expressiveness. Figure 4B is a spiking modulator. The work in Ref. [44] might be adapted to fit this classification; however, to the authors' knowledge, an ultrafast spiking modulator remains to be experimentally demonstrated. Figure 4C illustrates a purely electronic

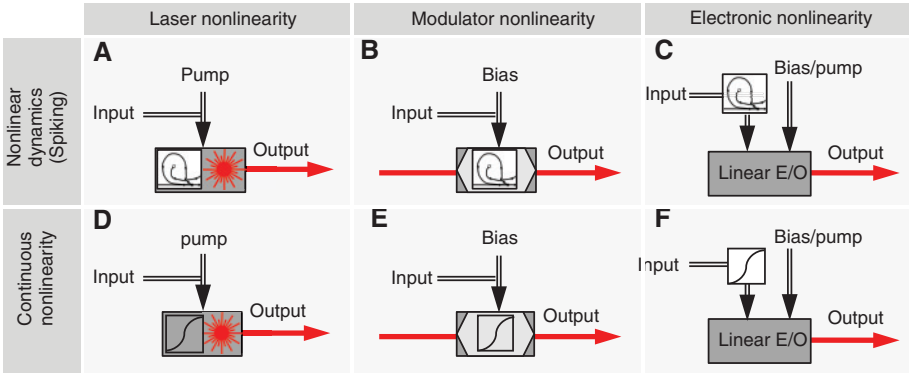

Figure 4: Classification of O/E/O PNN nonlinearities and possible implementations.
(A) Spiking laser neuron. (B) Spiking modulator. (C) Spiking or arbitrary electronic system driving a linear electro-optic (E/O) transducer – either modulator or laser. (D) Overdriven continuous laser neuron, as demonstrated in Ref. [41]. (E) Continuous modulator neuron, as demonstrated in Ref. [42]. (F) Continuous purely electronic nonlinearity with optical output. From Ref. [18].

approach to nonlinear neural behavior. Linear E/O could be done by either a modulator or a directly driven laser. This class could encompass interesting intersections with efficient analog electronic neurons in silicon [45, 46]. A limitation of these approaches is the need to operate slow enough to digitize outputs into a form suitable for electronic TDM and/or AER routing.

Figure 4D describes a laser with continuous nonlinearity, an instantiation of which was recently demonstrated in Ref. [41]. Figure 4E shows a modulator with continuous nonlinearity, the first demonstration of which in a PNN and recurrent network is presented in [42]. The pros and cons between the schemes in Figure 4D and E are the same ones brought up by the on-chip vs. off-chip light source debate, currently under way in the silicon photonics community. On-chip sources could provide significant energy savings [47]. They require the introduction of exotic materials to the silicon photonics process to provide optical gain, but active research in this area has the goal of making this feasible [48, 49]. The opposing school of thought argues that on-chip sources are still a nascent technology [50]. Whereas fiber-to-chip coupling presents practical issues [51], discrete laser sources are cheap and well understood. Furthermore, on-chip lasers dissipate large amounts of power [52], the full implications of which may complicate system design [50]. Modulator-based neurons could provide a more technologically feasible, although lower performing, alternative to spiking laser neurons for near-term large-scale integrated photonic neural systems. In either case, the conception of a PNN module, consisting of a photonic weight bank, detector, and E/O converter, as a participant in a broadcast-and-weight network could be applied to a broad array of neuron models and technological implementations.

Both discussed all-optical and O/E/O PNN approaches depend on charge carrier dynamics, whose lifetime eventually limits the bandwidth of the summation operation. The O/E/O strategy, however, has a few advantages: it can be modularized, it uses more standard optoelectronic components, and it is more amenable to integration. Therefore, here, we gave more attention to this strategy. Moreover, although the E/O part of the PNN can involve any kind of nonlinearity (Figure 4), not necessarily spiking, we are focusing on spiking behavior because of its interesting noise resistance and richness of representation. As such, we study here excitable semiconductor laser physics with the objective of directly producing optical spikes.

In this light, the PNN could be separated into three parts, just like the artificial neuron: weighting, addition, and neural behavior. Weighting and adding define how nonlinear nodes can be “networked” together, whereas the neural behavior dictates the “activation function” shown in Figure 2. In Section 3, we review the recent developments of semiconductor excitable lasers that emulate spiking neural behavior. In Section 4, we discuss a scalable WDM networking scheme.

3 PNN part I: excitable lasers

In the past few years, there has been a bloom of optoelectronic devices exhibiting excitable dynamics isomorphic to a physiological neuron. Excitable systems can be roughly defined by three criteria: (a) there is only one stable state at which the system can indefinitely stay at rest; (b) when excited above a certain threshold, the system undergoes a stereotypical excursion, emitting a “spike”; and (c) after the excursion, the system decays back to rest in the course

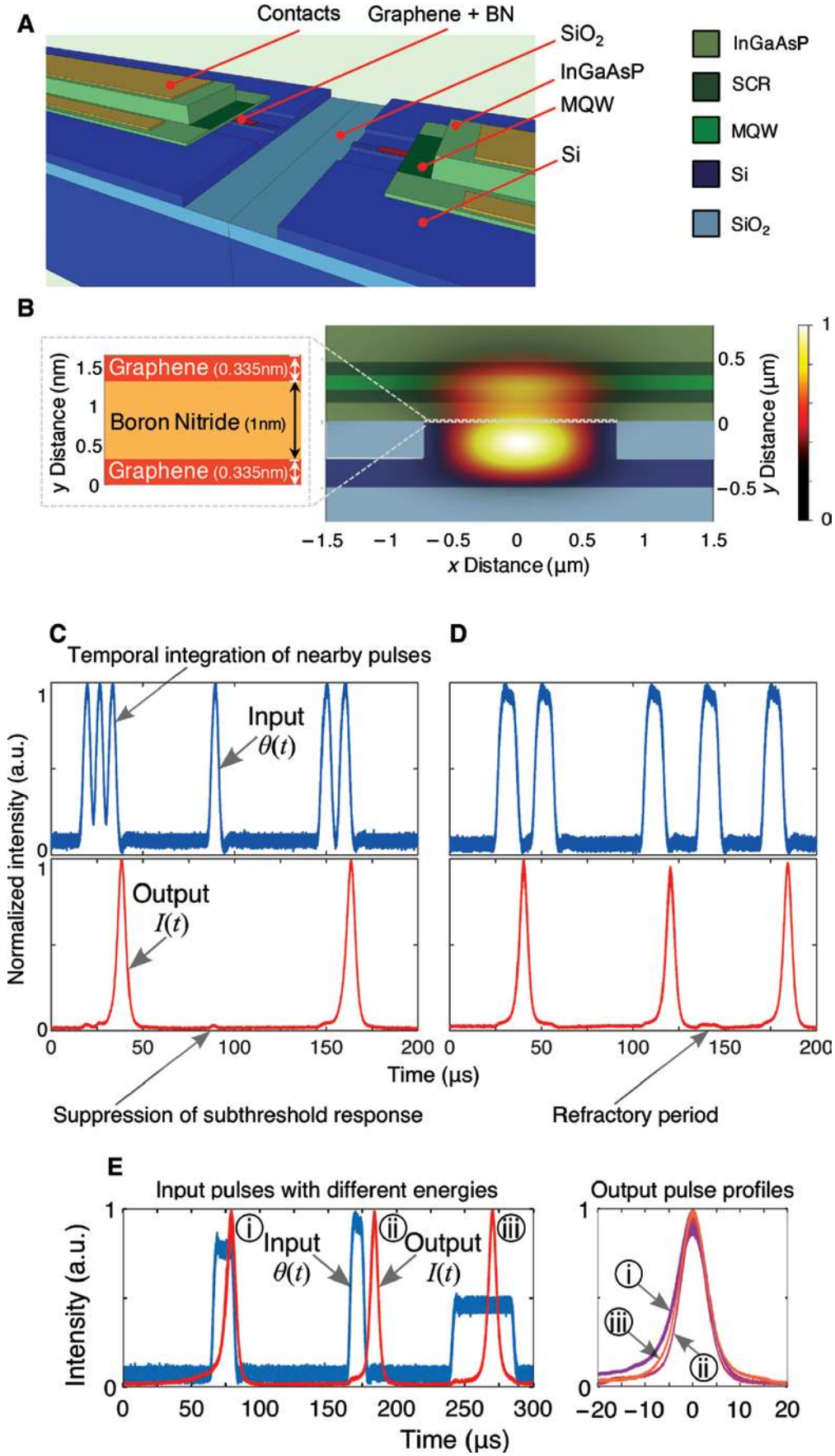
A
Contacts
Graphene + BN
SiO2
InGaAsP
MQW
Si
InGaAsP
SCR
MQW
Si
SiO2
B
y Distance (nm)
Graphene (0.335nm)
Boron Nitride (1nm)
Graphene (0.335nm)
x Distance (μm)
y Distance (μm)
C
Temporal integration of nearby pulses
Input
θ(t)
Output
I(t)
Normalized intensity (a.u.)
Time (μs)
Suppression of subthreshold response
D
Refractory period
E
Input pulses with different energies
Input
θ(t)
Output
I(t)
Intensity (a.u.)
Time (μs)
Output pulse profiles

Figure 5: Excitable dynamics of the graphene excitable laser.
Blue and red curves correspond to input and output pulses, respectively. (A) Cutaway architecture of a hybrid InGaAsP-graphene-silicon evanescent laser (not to scale) showing a terraced view of the center. (B) Cross-sectional profile of the excitable laser with an overlaid electric field (E-field) intensity $|\vec{E}|^2$ profile. (C–E) Excitable dynamics of the graphene "fiber" laser. (C) Excitatory activity (temporal integration of nearby pulses) can push the gain above the threshold, releasing spikes. Depending on the input signal, the system can have a suppressed response due to the presence of either subthreshold input energies (integrated power $\int|\theta(t)|^2\,dt$) or (D) a refractory period during which the laser is unable to pulse (regardless of excitation strength). (E) Restorative properties: repeatable pulse shape even when inputs have different energies. Reproduced from Shastri et al. [39]. Licensed under CC BY.

of a "refractory period" during which it is temporarily less likely to emit another spike.

3.1 Example of excitability behavior analogous to LIF

Excitable behavior can be realized near the threshold of a passively Q-switched two-section laser with saturable absorber (SA). Figure 5A and B shows an example of integrated design in a hybrid photonics platform. This device comprises a III-V epitaxial structure with multiple quantum well (MQW) region (the gain region) bonded to a low-loss silicon rib waveguide that rests on a silicon-on-insulator (SOI) substrate with sandwiched layers of graphene acting as an SA region with a sandwiched heterostructure of two monolayer graphene sheets and an hexagonal boron nitride (hBN) spacer. The gain section of this structure is electrically pumped. The full cavity structure includes III-V layers bonded to silicon and a quarter-shifted wavelength grating. The laser emits light along the waveguide structure into a passive silicon network. Figure 5C–E shows experimental data from a fiber ring laser prototype, demonstrating the key properties of excitability.

In general, the dynamics of a two-section laser composed of a gain section and an SA can be described by the Yamada model [Eqs. (2)–(4)] [53]. This 3D dynamical system, in its simplest form, can be described by the following undimensionalized equations [34, 37]:

$$\frac{dG(t)}{dt} = \gamma_G[A - G(t) - G(t)I(t)] + \theta(t) \tag{2}$$

$$\frac{dQ(t)}{dt} = \gamma_Q[B - Q(t) - aQ(t)I(t)] \tag{3}$$

$$\frac{dI(t)}{dt} = \gamma_I[G(t) - Q(t) - 1]I(t) + \epsilon f(G), \tag{4}$$

where $G(t)$ models the gain, $Q(t)$ is the absorption, $I(t)$ is the laser intensity, A is the bias current of the gain region, B is the level of absorption, a describes the differential absorption relative to the differential gain, γ_G is the relaxation rate of the gain, γ_Q is the relaxation rate of the absorber, γ_I is the inverse photon lifetime, $\theta(t)$ is the time-dependent input perturbations, and $\epsilon f(G)$ is the spontaneous noise contribution to intensity; ϵ is a small coefficient.

In simple terms, if we assume electrical pumping at the gain section, the input perturbations are integrated by the gain section according to Eq. (2). An SA effectively becomes transparent as the light intensity builds up in the cavity and bleaches its carriers. It was shown in [34] that the near-threshold dynamics of the laser described can be approximated to Eq. (5):

$$\frac{dG(t)}{dt} = -\gamma_G(G(t) - A) + \theta(t); \tag{5}$$

if $G(t) > G_{\text{thresh}}$, then release a pulse and set $G(t) \to G_{\text{reset}}$, where $G(t)$ models the gain, γ_G is the gain carrier relaxation rate, and A is the gain bias current. The input $\theta(t)$ can include spike inputs of the form $\theta(t) = \sum_i \delta_i(t - \tau_i)$ for spike firing times τ_i, G_{thresh} is the gain threshold, and $G_{\text{reset}} \sim 0$ is the gain at transparency.

$$C_m \frac{dV_m(t)}{dt} = -\frac{1}{R_m}(V_m(t) - V_L) + I_{\text{app}}(t);$$

if $V_m(t) > V_{\text{thresh}}$, then release a spike and set $V_m(t) \to V_{\text{reset}}$, where $V_m(t)$ is the membrane voltage, R_m is the membrane resistance, V_L is the equilibrium potential, and I_{app} is the applied current (input).

One can note the striking similarity to the LIF model in Eq. (1): setting the variables $\gamma_G = 1/R_mC_m$, $A = V_L$, $\theta(t) = I_{\text{app}}(t)/R_mC_m$, and $G(t) = V_m(t)$ shows their algebraic equivalence. Thus, the gain of the laser $G(t)$ can be thought of as a virtual "membrane voltage", the input current A as a virtual "equilibrium voltage", etc.

A remarkable difference can be observed between the two systems, though: whereas in the neural cell membrane the timescales are governed by an R_mC_m constant of the order of milliseconds, the carrier dynamics in lasers are as fast as nanoseconds. Although this form of excitability was found in two-section lasers, other device morphologies have also shown excitable dynamics. The advantage

Table 1: Characteristics of recent excitable laser devices. Note that this table does not have a one-to-one correspondence with Figure 4, because some of them are not E/O devices. However, we observed that devices A, D, and F belong to category 2.3.2(a) and device E resembles more closely category 2.3.2(c).

Device	Injection scheme	Pump	Excitable dynamics	Refs.
A. Two-section gain and SA	Electrical	Electrical	Stimulated emission	[34–37, 39, 54–61]
B. Semiconductor ring laser	Coherent optical	Electrical	Optical interference	[44, 62–65]
C. Microdisk laser	Coherent optical	Electrical	Optical interference	[32, 66]
D. 2D Photonic crystal nanocavity[a]	Electrical	Electrical	Thermal	[67–69]
E. Resonant tunneling diode PD and laser diode[b]	Electrical or incoherent optical	Electrical	Electrical tunneling	[70–72]
F. Injection-locked semiconductor laser with delayed feedback	Electrical	Electrical	Optical interference	[73–83]
G. Semiconductor lasers with optical feedback	Incoherent optical	Electrical	Stimulated emission	[84–90]
H. Polarization switching VCSELs	Coherent optical	Optical	Optical interference	[33, 91, 92]

[a]Technically, this device is not an excitable laser but an excitable cavity connected to a waveguide.
[b]The authors call it "excitable optoelectronic device" because the excitability mechanism lies entirely in an electronic circuit rather than the laser itself.

of constructing a clear abstraction to the LIF model is that it allows engineers to reuse the same methods developed in the computational neuroscience community for programming a neuromorphic processor. In the next section, we present recent optical devices with excitable dynamics.

3.2 Semiconductor excitable lasers

Optical excitability in semiconductor devices are being widely studied both theoretically and experimentally. These devices include multisection lasers, ring lasers, photonic crystal nanocavities, tunneling diode attached to laser diodes, and semiconductor lasers with feedback, as summarized in Table 1. We group them under the terminology "excitable lasers" for convenience, but exceptions are described in the caption of the table.

Generally speaking, these lasers use III-V quantum wells or quantum dots for efficient light generation. However, they fall into one of three injection categories (illustrated in Figure 3) and possess very diverse excitability mechanisms. It is difficult to group the rich dynamics of different lasers – which often requires a system of several coupled ordinary differential equations to represent it – using classification keywords. We focus on two fundamental characteristics: the way each laser can be modulated (injection scheme column) and on the physical effect that directly shapes the optical pulse (excitable dynamics column).

The injection scheme of the laser will determine whether it is compatible to all-optical PNNs (Section 2.3.1) or O/E/O PNNs (Section 2.3.2). Some of them (B, C, and H) operate free of electrical injection, meaning that bits of information remain elegantly encoded in optical carriers. However, as we have pointed out in Section 2.3, avoiding the E/O conversion is much more difficult when you are trying to build a weight-and-sum device compatible with WDM, which is an essential building block for scalable photonic neural networks (Section 4).

The excitable dynamics determines important properties such as energy efficiency, switching speed, and bandwidth of the nonlinear node. The "optical interference" mechanism typically means that there are two competing modes with a certain phase relationship that can undergo a 2π topological excursion and generating an optical pulse in amplitude at the output port. This mechanism is notably different from the others in which it does not require an exchange of energy between charge carrier populations and the cavity field. As a result, systems based on this effect are not limited by carrier lifetimes yet are vulnerable to phase noise accumulation. Other mechanisms include photon absorption, stimulated emission, thermo-optic effect, and electron tunneling. There, the electronic dynamics of the device governs the population of charge carriers available for stimulated emission, thereby dominating the timescale of the generated pulses. Models of these mechanisms and how they elicit excitability are comprehensively detailed in Ref. [93], but a quantitative comparison between performance metrics of lasers in Table 1 is still called for. Qualitatively, however, excitable lasers can simultaneously borrow key properties of electronic transistors, such as thresholding and cascadability (cf. Section 1.3).

In addition to individual laser excitability, there have been a few demonstrations of simple processing circuits. Temporal pattern recognition [39] and stable recurrent memory [39, 70, 74] are essential toy circuits that demonstrate the basic aspects of network compatibility.

3.3 Elemental circuits of excitable lasers

Although many neuromorphic semiconductor excitable lasers have been proposed and demonstrated, few have so far been interconnected in an integrated platform. In this section, we discuss simple circuits that could be constructed using only two excitable lasers and that could verify important properties of the tested technology. The first one tests the property of cascadability, which fundamentally demonstrates that the excitable dynamics can overcome noise and attenuation [44]. The second one tests the capability of pattern recognition, which fundamentally demonstrates the ability of such circuits to encode and decode information present in spike timing.

3.3.1 Cascadability

As discussed in Section 2, the concept of cascadability is crucial for creating strong recurrent connections and neural networks of more than one neuron. Recurrent connections are important in neuroscience because they enable attractor networks and short-term information retention, playing a crucial role in memory function and recall [94]. Cascadability also enables the propagation and multiplication of signals across the network, a necessary requirement for distributed processing.

Cascadability has been proposed and numerically demonstrated in both optically [44, 66] and electrically [39, 58, 60] injected lasers. Cascadability in optically injected PNNs presents a challenge because optical interference is sensitive to optical phase noise. On the contrary, in O/E/O PNNs, it presents a challenge because of the quantum efficiency limit – output pulses must contain more photons than the inputs required to trigger them. As discussed in Section 4, interconnection induces a power penalty to the optical signal's intensity. Therefore, to drive a scalably large number of PNNs, amplification could play a significant role in either the O/E or the E/O stage.

A stable recurrent circuit was prototyped in an excitable graphene fiber ring laser (Figure 6) [39]. This is a proof-of-concept demonstration of cascadability and pulse regeneration. This circuit represents a test of the device's ability to handle feedback and the stable shape of subsequent pulses is not only an indication of cascadability but also of signal fidelity restoration.

3.3.2 Temporal pattern recognition

In the context of neurobiology, networks of spiking neurons convert analog data (detected from the outside world) into spike trains and recognize spatiotemporal bit patterns. Spatiotemporal patterns play an important role in both visual [95] and audio [96] cortical processing. An interesting phenomenon that can happen in an SNN with fixed delays is "polychronization", as discovered by

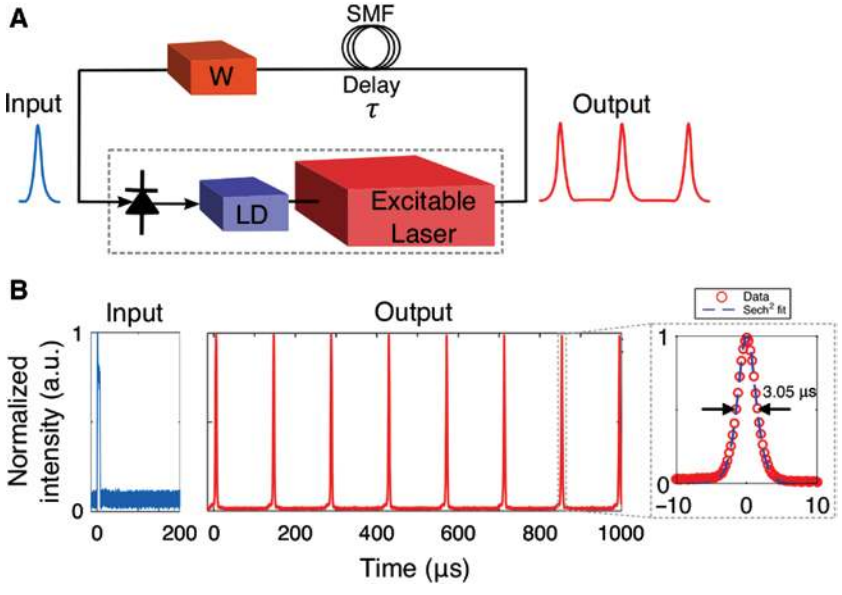

Figure 6: Self-recurrent bistable circuit.
(A) Set-up to test the self-referent connection. (B) Input and output waveforms. The first output pulse is fed back to the input after being delayed by ~100 μs, which initiates another excitatory pulse at the output. This recursive process results in a train of output pulses "ad eternum" at fixed intervals. Inset shows an output pulse profile and sech2 fitting curve. Reproduced from Shastri et al. [39]. Licensed under CC BY.

Izhikevich [97]. A subset of a large network of neurons can "polychronize" when a specific spatiotemporal stimulus is presented to a small number of neurons, and that triggers a repeatable, daisy-chain spiking pattern in the network. The neurons activated by the input pattern forms a "polychronous group". Thus, the polychronous group can recognize a particular spatiotemporal pattern input into a defined set of neurons. With synaptic plasticity, learning could occur due to strengthening, appearance, or extinction of polychronous groups, adding an elastic memory functionality to the network.

A simple pattern recognition circuit was prototyped by cascading two excitable graphene fiber ring lasers (Figure 7) with a delay τ between them. The objective was to distinguish (i.e. recognize) a specific input pattern: a pair of pulses separated by a time interval $\Delta t = \tau$, equal to the delay between the excitable lasers.

This simple circuit demonstrates important features necessary for robust optical processing: well-isolated input/output ports allow for the construction of feedforward networks, and the spatiotemporal recognition of spikes allows the system to classify patterns. We expect more complex recognition and decoding as the number of neurons is increased.

The recent progress in the field of integrated excitable lasers is very encouraging. We identify in the literature a collection of researchers in different parts of the world producing responsible, scholarly work founded in experimental validation and first principles. Today, multiproject wafer services offer rapid prototyping of concept systems in multiple platforms [98, 99], while device researchers are working toward a powerful and versatile active/passive photonic hybrid platform [19]. In addition, alternative implementations of a PNN offer flexibility with respect to which platform it could be instantiated. In summary, these advances together with thorough qualitative analysis have cleared the way for the creation of a reconfigurable photonic neuromorphic processor.

4 PNN part II: network architecture

4.1 Isomorphism to biological spiking neuron

Neurons only have computational capabilities if they are in a network. Therefore, an excitable laser (or spiking

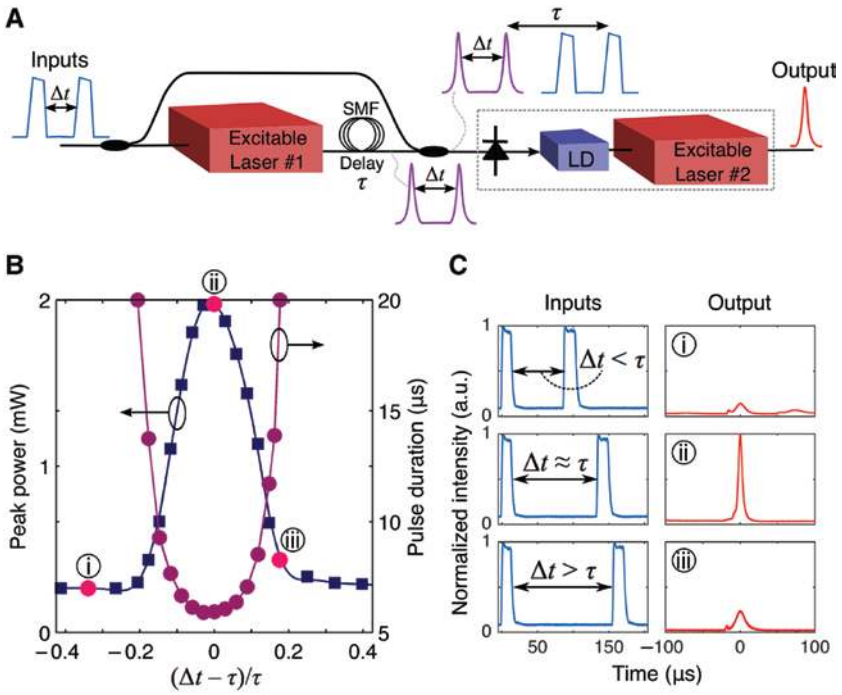

Figure 7: Temporal pattern recognition circuit.
(A) Simple circuit with two cascaded graphene excitable lasers. (B) Measured output pulse peak power and pulse duration as a function of the time interval between the two input pulses. (C) Measured input and output waveforms at specific instances: (i) $\Delta t - \tau = -45$ μs, (ii) $\Delta t \approx \tau = 135$ μs, and (iii) $\Delta t - \tau = 35$ μs. The output pulse energy is largest when $\Delta t = \tau$ showing the system only reacts to a specific spatiotemporal input pattern. Reproduced from Shastri et al. [39]. Licensed under CC BY.

laser) can only be viewed as a neuron candidate if it is contained in a PNN. The configurable analog connection strengths between neurons, called weights, are as important to the task of network processing as the dynamical behavior of individual elements. In Section 3, we have discussed several proposed excitable lasers exhibiting neural behavior and cascadability between these lasers. In this section, we discuss the challenges involving the creation of a network of neurons using photonic hardware, in particular, the creation of a weighted addition scheme for every PNN. Tait et al. [31] proposed an integrated photonic neural networking scheme called "broadcast-and-weight" that uses WDM to support a large number of reconfigurable analog connections using silicon photonic device technology.

A spiking and/or analog photonic network consists of three aspects: a protocol, a node that abides by that protocol (the PNN), and a network medium that supports multiple connections between these nodes. This section will begin with broadcast-and-weight as a WDM protocol in which many signals can coexist in a single waveguide and all nodes have access to all the signals. Configurable analog connections are supported by a novel device called a microring resonator (MRR) weight bank (Figure 8). Sections 4.3 and 4.4 summarize the experimental investigations of MRR weight banks.

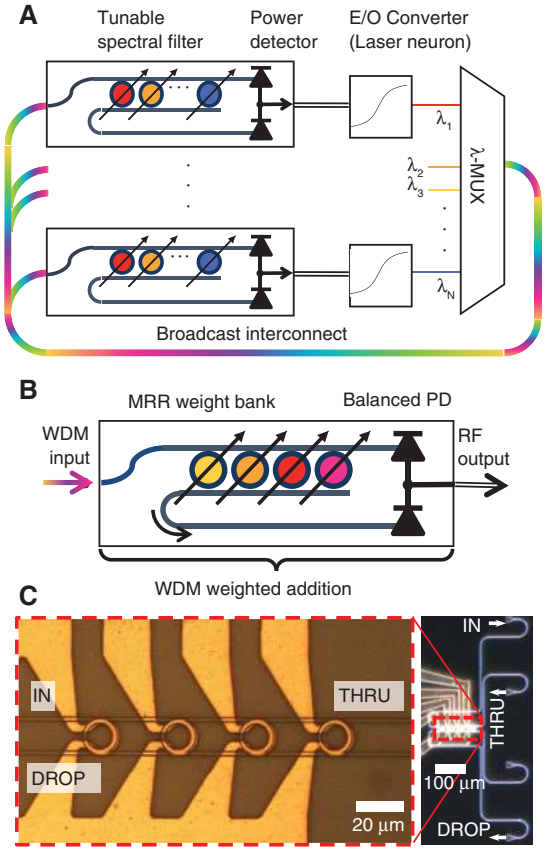

Figure 8: Configurable analog weights in neuromorphic photonics. (A) Broadcast-and-weight network. An array of source lasers outputs distinct wavelengths (represented by solid color). These channels are wavelength multiplexed (WDM) in a single waveguide (multicolor). Independent weighting functions are realized by tunable spectral filters 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 is transduced to an optical signal after nonlinear transformation. Adapted from Ref. [100]. (B) Tunable spectral filter constructed using MRR weight bank. Tuning MRRs between on- and off-resonance switches a continuous fraction of optical power between drop and through ports. A balanced PD yields the sum and difference of weighted signals. (C) Left: Optical micrograph of a silicon MRR weight bank, showing a bank of four thermally tuned MRRs. Right: Wide area micrograph, showing fiber-to-chip grating couplers [101]. Adapted from Ref. [102].

4.2 Broadcast-and-weight protocol

WDM channelization of the spectrum is one way to efficiently use the full capacity of a waveguide, which can have usable transmission windows up to 60 nm (7.5 THz bandwidth) [103]. In fiber communication networks, a WDM protocol called broadcast-and-"select" has been used for decades to create many potential connections between communication nodes [104]. In broadcast-and-select, 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. Broadcast-and-"weight" is similar but differs by directing multiple inputs simultaneously into each detector (Figure 8B) and with a continuous range of effective drop strengths between −1 and +1, corresponding to an analog weighting function.

The ability to control each connection, each weight, independently is a crucial aspect of neural network models. Weighting in a broadcast-and-weight network is accomplished by a tunable spectral filter bank at each node, an operation analogous to a neural weight. The local state of the filters defines the interconnectivity pattern of the network.

A great variety of possible weight profiles allows a group of functionally similar units to instantiate a tremendous variety of neural networks. A reconfigurable filter can be implemented by an MRR – in simple words, a waveguide bent back on itself to create an interference condition. The MRR resonance wavelength can be tuned thermally (as in Figure 8C) or electronically on timescales much slower than signal bandwidth. Practical, accurate, and scalable MRR control techniques are a critical step toward large-scale

analog processing networks based on MRR weight banks. We present them in Section 4.3. The analysis of scaling and design for MRR weight banks is then given in Section 4.4.

4.3 Controlling photonic weight banks

Sensitivity to fabrication variations, thermal fluctuations, and thermal crosstalk have made MRR control an important topic for WDM demultiplexers [105], high-order filters [106], modulators [107], and delay lines [108]. Commonly, the goal of MRR control is to track a particular point in the resonance relative to the signal carrier wavelength, such as its center or maximum slope point. On the contrary, an MRR weight must be biased at arbitrary points in the filter roll-off region to multiply an optical signal by a continuous range of weight values. Feedback control approaches are well suited to MRR demultiplexer and modulator control [109, 110], but these approaches rely on having a reference signal with consistent average power. In analog networks, signal activity can depend strongly on the weight values, so these signals cannot be used as references to estimate weight values. These reasons dictate a feedforward control approach for MRR weight banks.

4.3.1 Single-channel control accuracy and precision

How accurate can a weight be? The resolution required for effective weighting is a topic of debate within the neuromorphic electronics community, with IBM's TrueNorth selecting four digital bits plus one sign bit [111]. In Refs. [102, 112], the continuous weight control of an MRR weight bank channel was shown using an interpolation-based calibration approach. The goal of the calibration is to have a model of applied current/voltage vs. effective weight command. The calibration can be performed once per MRR and its parameters can be stored in memory. Once calibration is complete, the controller can navigate the MRR transfer function to apply the correct weight value for a given command. However, errors in the calibration, environmental fluctuations, or imprecise actuators cause the weight command to be inaccurate. It is necessary to quantify that accuracy.

Analog weight control accuracy can be characterized in terms of the ratio of weight range (normalized to 1.0) to worst-case weight inaccuracy over a sweep and stated in terms of bits or a dynamic range. The initial demonstration reported in Ref. [102] indicates a dynamic range of the weight controller of 9.2 dB – in other words, an equivalent digital resolution of 3.1 bits.

4.3.2 Multichannel control accuracy and precision

Another crucial feature of an MRR weight bank is the simultaneous control of all channels. When sources of crosstalk between one weight and another are considered, it is impossible to interpolate the transfer function of each channel independently. Simply extending the single-channel interpolation-based approach of measuring a set of weights over the full range would require a number of calibration measurements that scale exponentially with the channel count, as the dimension of the range grows with channel count. Simultaneous control in the presence of crosstalk therefore motivates model-based calibration approaches.

Model-based, as opposed to interpolation-based, calibration involves parameterized models for crosstalk-inducing effects. The predominant sources of crosstalk are thermal leakage between nearby integrated heaters and, in a lab set-up, interchannel cross-gain saturation in fiber amplifiers, although optical amplifiers are not a concern for fully integrated systems that do not have fiber-to-chip coupling losses. Thermal crosstalk occurs when the heat generated at a particular heater affects the temperature of neighboring devices (see Figure 8C). In principle, the neighboring channel could counter this effect by slightly reducing the amount of heat its heater generates. A calibration model for thermal effects provides two basic functions: forward modeling (given a vector of applied currents, what will the vector of resultant temperatures be?) and reverse modeling (given a desired vector of temperatures, what currents should be applied?). Models such as this must be calibrated to physical devices by fitting parameters to measurements. Calibrating a parameterized model requires at least as many measurements as free parameters. Ref. [113] described a method for fitting parameters with $O(N)$ spectral and oscilloscope measurements, where N is the number of MRRs. As an example, whereas an interpolation-only approach with 20 points resolution would require $20^4 = 160{,}000$ calibration measurements, the presented calibration routine takes roughly $4 \times [10(\text{heater}) + 20(\text{filter}) + 4(\text{amplifier})] = 136$ total calibration measurements. Initial demonstrations achieved simultaneous four-channel MRR weight control with an accuracy of 3.8 bits and precision of 4.0 bits (plus 1.0 sign bit) on each channel (Figure 9). Although optimal weight resolution is still a topic of discussion in the neuromorphic electronics community [9], several state-of-the-art architectures with dedicated weight hardware have settled on 4-bit resolution [111, 115].

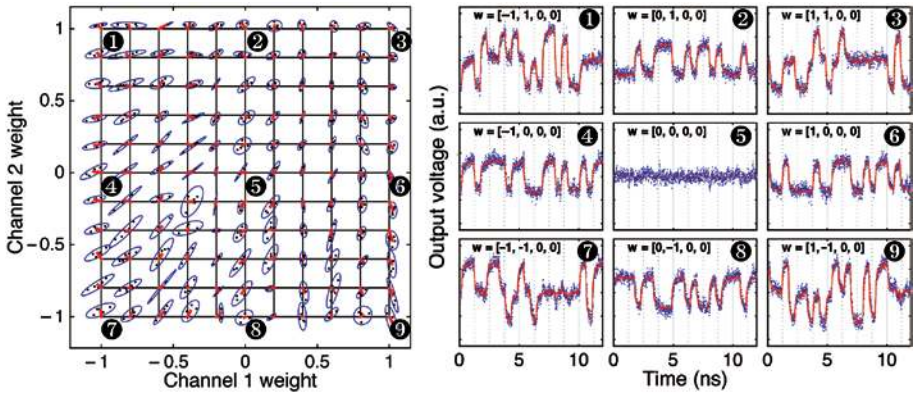

Figure 9: Demonstration and characterization of multi-channel analog weight control using microring resonator weight banks shown in Fig. 8. (A) 2D weight sweep showing controller accuracy and precision. After the calibration procedure, the target weight was swept five times over a grid of values from −1 to 1 (black grid). Black points are measured weight data. Red lines show the mean offset from each target grid point. Blue ellipses indicate one standard deviation around the mean. From this plot, it is deduced that the weight can be controlled with an accuracy of 3.8 bits. (B) [6, 14, 32, 37, 51, 73, 84, 111, 114] Output time trace of signals corresponding to points labeled in (A). The expected weighted signal is in red, whereas measured traces are in blue. From Ref. [113].

4.4 Quantitative analysis for photonic weight banks

Engineering analysis and design rely on quantifiable descriptions of performance called metrics. The natural questions of "how many channels are possible" and subsequently "how many more or fewer channels are garnered by a different design" are typically resolved by studying trade-offs. Increasing the channel count performance metric will eventually degrade some other aspects of performance until the minimum specification is violated.

Studying trade-offs between these metrics are important for better designing the network and understanding its limitations. Just as the case with control methodologies, it was found that quantitative analysis for MRR weight banks must follow an approach significantly different from those developed for MRR demultiplexers and modulators [100].

In conventional analyses of MRR devices for multiplexing, demultiplexing, and modulating WDM signals, the trade-off that limits channel spacing is interchannel crosstalk [103, 116, 117]. However, unlike MRR demultiplexers where each channel is coupled to a distinct waveguide output [105], MRR weight banks have only two outputs with some portion of every channel coupled to each. All channels are meant to be sent to both detectors in some proportion, so the notion of crosstalk between signals breaks down (Figure 8B). Instead, for dense channel spacing, different filter peaks viewed from the common drop port begin to merge together. This has the effect of reducing the weight bank's ability to weigh neighboring signals independently. To quantify this effect as a power penalty, the cross-weight penalty metric must include a notion of tuning "range" (Section 4.4.1). After this has been described, an example channel density analysis is carried out to derive the scalability of weight banks that use microresonators of a particular finesse (Section 4.4.2).

4.4.1 Cross-weight power penalty metric

In the single-channel case, an ideal tunable weight bank possesses a range of tuning states that include directing an incident optical signal completely to a through port (positive weight), completely to a drop port (negative weight), or to any intermediate ratio of both (Figure 8B). If a real weight incurs some loss, its weight range becomes a subset of the ideal. If there is a difference in loss between the drop and through ports, then the attainable weight range will also be unbalanced. Because the neural network abstraction should be able to provide a programmer with a range of weights from −1 to +1, we require that the range is usable only up to the minimum absolute extremum. Comparing the usable range to the ideal range yields a ratio, W, which quantifies the real device's ability to perform tunable optical weighting.

$$cW(1-\mathrm{D}) = \min\left[\max_{p}(\mu),\ \max_{p}(-\mu)\right], \tag{6}$$

where p is the tuning parameter and μ is the weight.

In the N-channel case, the ideal WDM weight bank is able to switch WDM channels completely independently from one another. However, if a given tuning parameter can affect multiple weight values, then the bank's weight range cannot be linearly separated into a composition of nonideal single-channel weight ranges. In other words,

the N-dimensional range of states becomes warped. Figure 10 depicts this mapping for a simulated two-channel bank that is parameterized by the MRR detunings.

As in the 1D case, a usable range can be defined as the largest balanced interval (i.e. a zero-centered square in 2D) that is completely covered by the attainable weight range. The usable range (green square in Figure 10B) is compared to the theoretical ideal (black bounding box in Figure 10B) to obtain an amplitude ratio between usable and ideal – a fill factor W.

This definition of cross-weight penalty can be extended conceptually to higher dimensions and WDM weight banks with an arbitrary number of channels. In N dimensions, the boundary is a $(N-1)$-dimensional closed manifold parameterized by the $(N-1)$-dimensional vector $\vec{s}$. The cross-weight penalty can then be defined as

$$cW_x(\mathrm{N-D}) = \min_{\vec{s}}\left[\max_{i\in 1\ldots N}|\mu_i(\vec{s})|\right]. \quad (7)$$

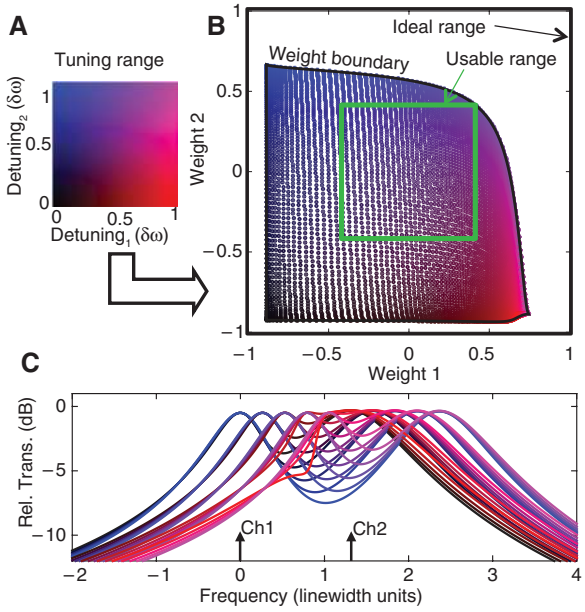

Figure 10: Example of cross-weight power penalty in a two-channel MRR weight bank.
(A) The device has two tuning degrees-of-freedom, which are resonance detunings of each filter. A (red, blue) color vector is used to indicate tuning state, which means that (A) depicts (red = x, blue = y). (B) The range of possible weight states attainable by the weight bank relative to the ideal range (outer bounding box). (red, blue) color indicates the tuning state that maps to a particular weight point. The usable range (green box) is graphically the largest square that lies fully within the possible weight range centered at zero. (C) Drop port spectra of the same model over a 5×5 parameter grid, with trace color used to indicate tuning. Frequency is normalized so that the MRR 1 peak has a center of 0 and full-width half-maximum (FWHM) of 1.0. Channel spacing in this simulation is 1.31 line widths and waveguide loss is 2 dB cm^{-1}. From Ref. [100].

W_x quantifies the "effective insertion loss" of a photonic weight bank, provided that it is capable of fully independent and balanced control. Supposing $W_x = 0.5$, then the weight bank is equivalent to an ideal $W_x = 1.0$ weight bank with an insertion loss of 0.5. W_x can therefore be stated as a power penalty in dB: $-10\log(W_x)$ describes the additional input power (in dB) required to make a non-ideal weight bank behave as an ideal weight bank.

4.4.2 Weight bank channel limits

The final step of channel density analysis is to study the degradation of a limiting metric as WDM channel spacing becomes more dense. A useful figure of merit for discussing the efficacy of a resonator-based circuit at a WDM task is the ratio of finesse to channel count. A theoretical minimum of this figure is 1.0.

Figure 11 shows the resulting power penalty contours of $-10\log(W_x)$ vs. channel spacing, δω, and bus length

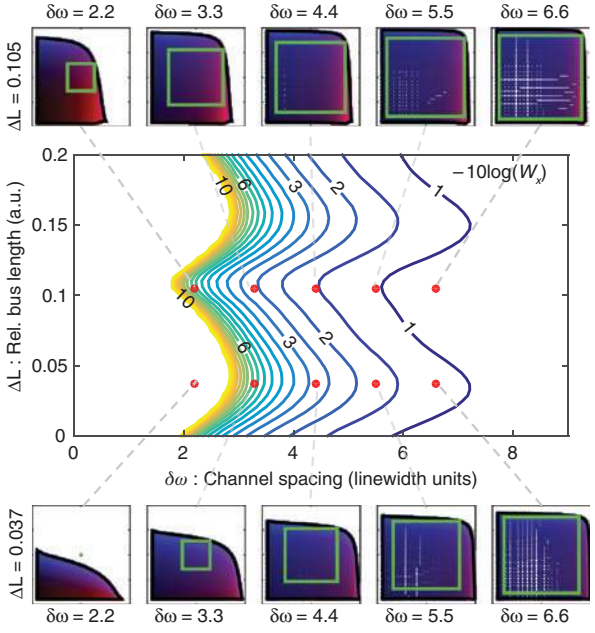

Figure 11: Cross-weight power penalty surface as a function of channel spacing δω and bus WG length offset ΔL.
Power penalty contours are plotted at 0.5 dB increments between 1 dB (blue) and 10 dB (yellow). The penalty increases as channel density decreases, eventually reaching an asymptote. This trade-off also depends significantly and approximately periodically on ΔL, indicating the influence of coherent multi-MRR interactions in the bus WGs. (Outer panels) Ranges of possible weight states, plotted as in Figure 10B, at 10 selected operating points that are indicated in Figure 10A by red circles. The top row, $\Delta L = 0.105$, represents the best-case trade-off between power and channel density, and the bottom row, $\Delta L = 0.037$, represents the worst-case. From Ref. [100].

changes, ΔL. The penalty is asymptotic in channel spacing, meaning there is an absolute minimum channel spacing regardless of acceptable power penalty. The power penalty cannot quite reach 0 dB because of optical losses. In Ref. [100], Tait et al. discovered that both the channel density wall and the trade-off between density and power are significantly affected by bus length changes. The resulting approximate periodicity (here, ~0.12 in arbitrary length units) is indicative of a coherent multi-MRR interference condition that could be exploited to decrease the power penalty figure [100]. What's perhaps surprising is that the effect of bus length remains significant even when channels are spaced relatively far apart. The 1 dB contour line (blue) fluctuates between 2.7 and 3.4 line widths over a period of ΔL.

WDM channel spacing, $\delta\omega$, can be used to determine the maximum channel count given a resonator finesse. Whereas finesse can vary significantly with the resonator type, normalized spacing is a property of the circuit (i.e. multiplexer vs. modulators vs. weight bank). Making an assumption that a 3 dB cross-weight penalty is allowed, we find that the minimum channel spacing falls between 3.41 and 4.61 line widths depending on bus length. High finesse silicon MRRs, such as that shown in Refs. [118] (finesse = 368) and [119] (finesse = 540), could support 108 and 148 channels, respectively. Other types of resonators in silicon, such as elliptical microdisks [120] (finesse = 440) and traveling-wave microresonators [121] (finesse = 1140), could reach up to 129 and 334 channels, respectively.

MRR weight banks are an important component of neuromorphic photonics – regardless of PNN implementation – because they control the configuration of analog network linking photonic neurons together. In Ref. [113], it was concluded that ADC resolution, sensitivized by biasing conditions, limited the attainable weight accuracy. Controller accuracy is expected to improve by reducing the mismatch between tuning range of interest and driver range. Ref. [100] arrived at a scaling limit of 148 channels for an MRR weight bank, which is not impressive in the context of neural networks. However, the number of neurons could be extended beyond this limit using spectrum reuse strategies (Figure 12) proposed in Ref. [31] by tailoring interference within MRR weight banks as discussed in [100] or by packing more dimensions of multiplexing within silicon waveguides, such as mode-division multiplexing. As the modeling requirements for controlling MRR weight banks become more

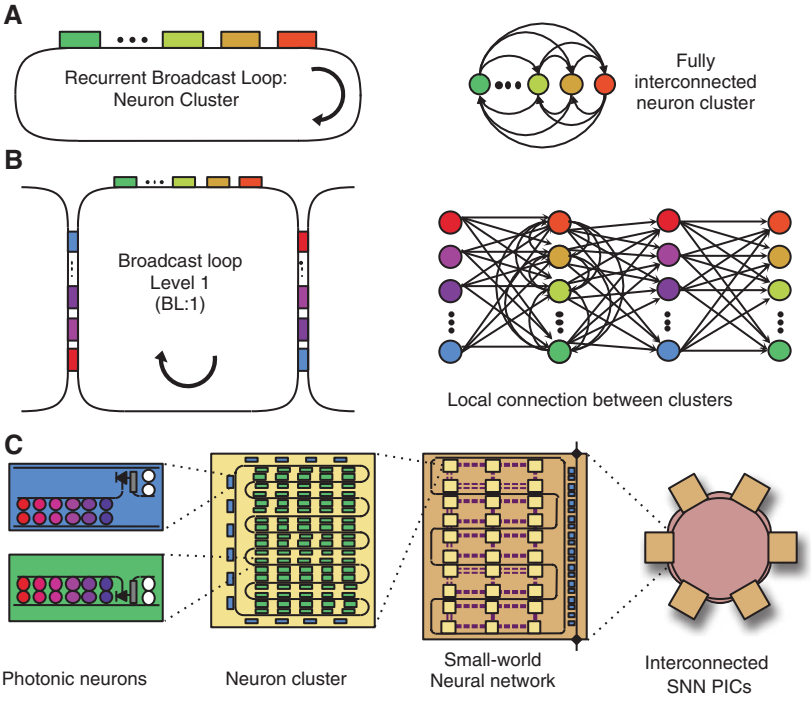

Figure 12: Spectrum reuse strategy.
(A) Fully interconnected network by attaching PNNs to a broadcast loop (BL) waveguide. (B) Slightly modified PNN can transfer information from one BL to another. (C) Using this scheme, neuron count in one chip is only limited by footprint, but PICs can be further interconnected in an optical fiber network.

computationally intensive, a feedback control technique would be transformative for both precision and modeling demands. Despite the special requirements of photonic weight bank devices making them different from communication-related MRR devices, future research could enable schemes for feedback control.

5 Neuromorphic platform comparison

We have recently produced a quantitative comparison between neuromorphic hardware architectures [18, 20]. Weighted addition is critical for neural network implementations, and as the number of operations scales quadratically with the number of nodes in all-to-all connected networks, it represents the most costly hardware scalability bottleneck [9]. Thus, for analysis, we can deconstruct this operation as a parallelized set of MACs and use it as a reference unit of computation. The MAC operation takes the following form: $a \leftarrow a + (w + x)$. It includes both a multiplication (i.e. x is multiplied by the "weight" w) and an addition (the result is accumulated to variable a).

For consistency, we compare architectures that have similar functionality: we limit ourselves to fully reconfigurable systems of SNNs. The analysis includes electronic neuromorphic architectures introduced in Section 1.1. For the photonically enhanced system, we studied an optoelectronic neural network with PNNs instantiated within the hybrid silicon/III-V platform [58, 122]. We also consider a future photonic crystal instantiation based on fundamental physical considerations. Calculated metrics are based on realistic device parameters derived from the literature.

Results are summarized in Table 2. The most striking figure is the number of operations per second, which exceeds electronic platforms by three orders of magnitude compared to the analog/digital accelerated HICANN and three orders of magnitude compared to the others that are purely digital implementations. This stems from both the high bandwidths and low latencies possible with photonic signals. The optoelectronic approach is also able to achieve energy efficiencies that are on the same order of magnitude as those in electronics, which avoids the heat problems that have prevented digital CMOS electronics from reaching similar operating bandwidths. The optoelectronic approach is able to achieve such energy efficiency at high speeds because power is mainly consumed statically by the lasers, whereas the passive filters have low leakage current. This contrasts to CMOS digital switches, whose power consumption increases dynamically with clock speed. Processor fan-in is similar in both platforms despite very differing technologies. The area per MAC is more stringent in a photonically enhanced system, as photonic elements cannot be shrunk beyond the diffraction limit of light. This is because each data channel requires a weighting filter in the PNN, such as an MRR pair, which adds a footprint penalty. However, this is compensated by the fact that a single waveguide can carry many wideband channels simultaneously, unlike electronic wires. Nonetheless, although photonically enhanced systems cannot compete with the miniaturization of future nanoelectronics, the estimated footprint of such a system is currently on par with some of the electronic systems presented here.

Table 2: Comparison between different neuromorphic processors.

Chip	MAC rate per processor	Energy per MAC (pJ)	Processor fan-in	Area per MAC (μm^2)	Synapse precision (bit)
Photonic hybrid III-V/Si (current work)	20 GHz	1.3	108	205	5.1
Sub-λ photonics (future trend)	200 GHz	0.0007	~200	20	8
HICANN [12]	22.4 MHz	198.4	224	780	4
TrueNorth [11]	2.5 kHz	0.27	256	4.9	5
Neurogrid [10]	40.1 kHz	119	4096	7.1	13
SpiNNaker[a] [13]	3.2 kHz	6e5	320	217	16

III-V/Si hybrid stands for estimated metrics of an SNN in a PIC in a III-V/Si hybrid platform. Sub-λ stands for estimated metrics for a platform using optimized subwavelength structures, such as photonic crystals. An MAC event occurs each time a spike is integrated by the neuron. Neuron fan-in refers to the number of possible connections to a single neuron. The energy per MAC for HICANN, TrueNorth, Neurogrid, and SpiNNaker was estimated by dividing wall-plug power to number of neurons and to operational MAC rate per processor. The area per MAC was estimated by dividing the chip/board size to the number of MAC units (neuron count times fan-in). All numbers therefore include overheads in terms of footprint and area.

[a]Neurons, synapses, and spikes are digitally encoded in event headers that travel around cointegrated processor cores. Therefore, all numbers here are based on a typical application example.

6 Outlook

After half a century of continuous investment and commercial success, digital CMOS electronics dominates the industry of general-purpose computing. However, with growing demand for connectivity, there is an urgent need for ultrafast coprocessors that could relieve the stress in digital processing circuits. Here, we have presented the elements of a reconfigurable photonic hardware that can emulate SNNs operating a billion times faster than the brain. As we identify proper metrics for a neuromorphic photonic processor, research efforts are incipiently transitioning from individual devices to systems design. We are witnessing a fast maturation of standardized photonic foundries in several platforms. Chrostowski and Hochberg [98] said that we are entering a nascent era of fabless photonics, where users can create computer-assisted chip designs and have it fabricated by these foundries using quality-controlled repeatable processes. We anticipate that neuromorphic photonic coprocessors (Figure 13) will be fabricated and packaged using fabless services in the medium term.

Applications for neuromorphic photonic processors can be clustered into two categories: (1) a front-end stage for radiofrequency (RF) systems and data centers and (2) ultrafast processing for specialized fast applications [18]. The first category uses the low-latency, parallelism, and energy-efficient properties of photonics to alleviate the throughput of RF systems, for example, by executing dimensionality reduction tasks such as principal component analysis or blind-source separation. The second category takes advantage of the raw speed (bandwidth and latency) of the photonic processor to execute iterative algorithms mapped to recurrent neural networks.

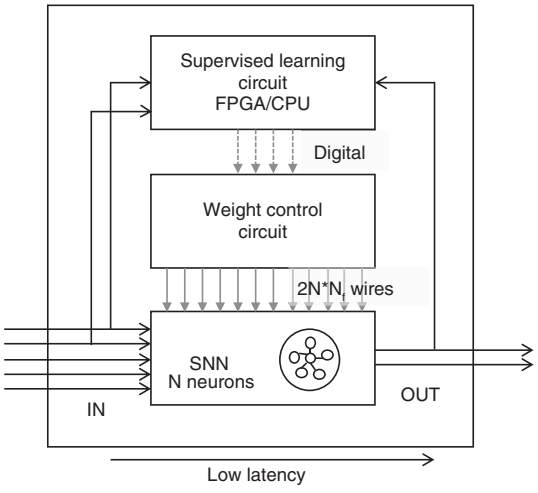

Figure 13: Diagram description of a fully packaged neuromorphic processor.
Whereas two layers of electronics provide reconfigurability, the photonic SNN permits low-latency functionality. N_i: Fan-in of each neuron.

Neuromorphic photonic processors join a class of photonic hardware accelerators designed to assist in acquisition, feature extraction, and storage of wideband waveforms [123]. These accelerators manipulate the spectrotemporal of a wideband signal, a task difficult to accomplish in analog electronics over broad bandwidth and with low loss. Reservoir computing is another promising model of analog computing. In reservoir-based models, a fixed complex system (the reservoir) generates an enormous number of nonlinear functions of inputs, and then a readout layer is trained to approximate the desired task out of a linear combination of reservoir functions. Reservoir computers consisting of a photonic reservoir with electronic readout layer have received substantial recent attention from the photonics community and have experimentally demonstrated a range of machine learning tasks [124–128].

6.1 Real-time RF processing

After some initial front-end processing (i.e. heterodyning and amplification), most radio transceiver systems are processed by either DSPs or field programmable gate arrays (FGPAs) for more complex signal operations. However, the speeds of these processors (i.e. ~500 MHz) limit the overall throughput of RF carrier signals, which can easily be in GHz range. Clever sampling and parallelization can help alleviate this bottleneck but at the cost of much higher latency and a significant resource/energy overhead. Specialized RF application-specific integrated circuits (ASICs) are another option but are expensive, require significant development time, and have limited reconfigurability. Future imagined multiple-in multiple-out (MIMO) systems – which, in the case of massive MIMO, can be on the order of ~100 s of input and output channels [129, 130] – are especially susceptible to this bottleneck and may require a radically new solution.

Adding a photonic processing chip to the front of a radio transceiver would allow very complex operations to be performed in real time, which can significantly offload electronic postprocessing and provide a technology to make faster, more relevant RF decisions on-the-fly. Massive MIMO systems based on beamforming in phased array antennas require a processor that can distinguish and operate on hundreds of high bandwidth signals simultaneously, a feat that is currently speed limited by

current electronic processors [129, 131]. A photonic neural network model is a perfect fit for addressing this kind of technological challenge: efficient MIMO beamforming relies on MAC operations that are already applied in neural network models via "weighted addition". In addition, classification algorithms can be built efficiently using the neural network approach, allowing for RF fingerprinting and signal identification.

6.2 Nonlinear programming

Another way of taking advantage of raw speed is via an "iterative" approach. Iterative algorithms find successfully better approximations to a problem of interest and often require many time steps to reach a desired solution. A large class of problems that can be solved iteratively include "linear and nonlinear programming problems".

Quadratic programming (QP) are optimization problems with quadratic multivariable objective function subject to constraints. A notable example of a computational problem that can be reduced to a QP includes model predictive control (MPC). The ability of MPC to handle large MIMO systems with physical constraints has led to very successful applications in slow processes, where there is sufficient time for solving the optimization problem between sampling instants. The application of MPC to faster systems, therefore, relies on new ways of finding faster solutions to QP problems [132]. It has been shown that QPs can be mapped onto recurrent neural networks that converge to an attractor state corresponding to the solution of QPs [133].

Because one of the most salient advantages of a photonic approach is its low time-of-flight (in picoseconds) between communicating processors, the convergence rates can be significantly improved by implementing them on a photonic platform. These processors represent some of the most effective yet generalized tools for acquiring and processing information and controlling highly mobile systems, such as a hypersonic aircraft [134].

Acknowledgments: This work was supported in part by the DARPA under contract no. DARPA-BAA-12-64. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of DARPA or the U.S. Government. 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 (NSERC). M.A.N. and A.N.T. acknowledge the support of the National Science Foundation Graduate Research Fellowship Program (NSF GRFP). The fabrication support for MRR weight banks was provided via the NSERC Silicon Electronic-Photonic Integrated Circuits (SiEPIC) Program and Richard Bojko at the University of Washington Nanofabrication Facility, part of the NSF National Nanotechnology Infrastructure Network (NNIN).

References

[1] Kim NS, Austin T, Baauw D, et al. Leakage current: Moore's law meets static power. Computer 2003;36:68–75.
[2] Dennard R, Gaensslen F, Yu W-N, Rideout L, Bassous E, Le Blanc A. Design of ion-implanted MOSFET's with very small physical dimensions. IEEE J Solid State Circuits 1974;9:257–68.
[3] Esmaeilzadeh H, Blem E, St. Amant R, Sankaralingam K, Burger D. Dark silicon and the end of multicore scaling. IEEE Micro 2012;32:122–34.
[4] Miller DAB. Attojoule optoelectronics for low-energy information processing and communications: a tutorial review, 2016.
[5] Taylor MB. Is dark silicon useful? Harnessing the four horsemen of the coming dark silicon apocalypse. In: Proceedings of Design Automation Conference 2012:1131–6.
[6] Andrae A, Edler T. On global electricity usage of communication technology: trends to 2030. Challenges 2015;6:117–57.
[7] Kachris C, Tomkos I. A survey on optical interconnects for data centers. IEEE Commun Surv Tutorials 2012;14:1021–36.
[8] Hochberg M, Harris NC, Ding R, et al. Silicon photonics: the next fabless semiconductor industry. IEEE Solid State Circuits Mag 2013;5:48–58.
[9] Hasler J, Marr B. Finding a roadmap to achieve large neuromorphic hardware systems. Front Neurosci 2013;7:118.
[10] Benjamin B, Gao P, McQuinn E, et al. Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations. Proc IEEE 2014;102:699–716.
[11] Merolla PA, Arthur JV, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science 2014;345:668–73.
[12] Schemmel J, Briiderle D, Griibl A, Hock M, Meier K, Millner S. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems. IEEE, 2010:1947–50.
[13] Furber S, Galluppi F, Temple S, Plana L. The SpiNNaker project. Proc IEEE 2014;102:652–65.
[14] The HBP Report. Technical report, The Human Brain Project, 2012.
[15] Miller DAB. Rationale and challenges for optical interconnects to electronic chips. Proc IEEE 2000;88:728–49.
[16] Boahen K. Point-to-point connectivity between neuromorphic chips using address events. Circuits Syst II Analog Digital Signal Process IEEE Trans 2000;47:416–34.
[17] Keyes RW. Optical logic-in the light of computer technology. Opt Acta Int J Opt 1985;32:525–35.
[18] Prucnal PR, Shastri BJ, Tait AN, Nahmias MA, Ferreira de Lima T. Neuromorphic photonics. CRC Press, Boca Raton, FL, USA, 2017.

[19] Liang D, Roelkens G, Baets R, Bowers JE. Hybrid integrated platforms for silicon photonics. Materials 2010;3:1782.
[20] Nahmias MA, de Lima TF, Tait AN, Shastri BJ, Prucnal PR. Photonically-enhanced neural networks: technology comparison. In: IEEE Photonics Conference. In preparation.
[21] Fok MP, Deming H, Nahmias M, Rafidi N, Rosenbluth D, Tait A, Tian Y, Prucnal PR. Signal feature recognition based on lightwave neuromorphic signal processing. Opt Lett 2011;36:19–21.
[22] Kravtsov KS, Fok MP, Prucnal PR, Rosenbluth D. Ultrafast all-optical implementation of a leaky integrate-and-fire neuron. Opt Express 2011;19:2133–47.
[23] Rosenbluth D, Kravtsov K, Fok MP, Prucnal PR. A high performance photonic pulse processing device. Opt Express 2009;17:22767–72.
[24] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35:1798–828.
[25] Maass W. Networks of spiking neurons: the third generation of neural network models. Neural Netw 1997;10:1659–71.
[26] Eliasmith C. A unified approach to building and controlling spiking attractor networks. Neural Comput 2005;17:1276–314.
[27] Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 1982;79:2554–8.
[28] Perrett DI, Rolls ET, Caan W. Visual neurones responsive to faces in the monkey temporal cortex. Exp Brain Res 1982;47:329–42.
[29] Thorpe S, Delorme A, Rullen RV. Spike-based strategies for rapid processing. Neural Netw 2001;14:715–25.
[30] Izhikevich EM. Which model to use for cortical spiking neurons? IEEE Trans Neural Netw 2004;15:1063–70.
[31] Tait AN, Nahmias MA, Shastri BJ, Prucnal PR. Broadcast and weight: an integrated network for scalable photonic spike processing. J Lightw Technol 2014;32:3427–39.
[32] Alexander K, Van Vaerenbergh T, Fiers M, Mechet P, Dambre J, Bienstman P. Excitability in optically injected microdisk lasers with phase controlled excitatory and inhibitory response. Opt Express 2013;21:26182.
[33] Hurtado A, Javaloyes J. Controllable spiking patterns in long-wavelength vertical cavity surface emitting lasers for neuromorphic photonics systems. Appl Phys Lett 2015;107:241103.
[34] Nahmias MA, Shastri BJ, Tait AN, Prucnal PR. A leaky integrate-and-fire laser neuron for ultrafast cognitive computing. IEEE J Select Top Quantum Electron 2013;19:1–12.
[35] Selmi F, Braive R, Beaudoin G, Sagnes I, Kuszelewicz R, Barbay S. Relative refractory period in an excitable semiconductor laser. Phys Rev Lett 2014;112:183902.
[36] Selmi F, Braive R, Beaudoin G, Sagnes I, Kuszelewicz R, Barbay S. Temporal summation in a neuromimetic micropillar laser. Opt Lett 2015;40:5690–3.
[37] Barbay S, Kuszelewicz R, Yacomotti AM. Excitability in a semiconductor laser with saturable absorber. Opt Lett 2011;36:4476–8.
[38] Shastri B, Tait A, Nahmias M, Wu B, Prucnal P. Spatiotemporal pattern recognition with cascadable graphene excitable lasers. In: Photonics Conference (IPC), 2014 IEEE, 2014:573–4.
[39] Shastri BJ, Nahmias MA, Tait AN, Rodriguez AW, Wu B, Prucnal PR. Spike processing with a graphene excitable laser. Sci Rep 2016;6:19126.
[40] Shastri BJ, Tait AN, Nahmias M, Wu B, Prucnal P. Coincidence detection with graphene excitable laser. In: CLEO. Optical Society of America, 2014:STu3I.5.
[41] Nahmias MA, Tait AN, Tolias L, et al. An integrated analog O/E/O link for multi-channel laser neurons. Appl Phys Lett 2016;108:151106.
[42] Tait A, Wu A, Zhou E, et al. Demonstration of a silicon photonic neural network. In: Summer Topicals Meeting Series (SUM). IEEE, 2016.
[43] Izhikevich EM. Dynamical systems in neuroscience: the geometry of excitability and bursting. Vol. 25. MIT Press, Cambridge, MA, USA, 2006.
[44] Van Vaerenbergh T, Fiers M, Mechet P, et al. Cascadable excitability in microrings. Opt Express 2012;20:20292.
[45] Indiveri G, Linares-Barranco B, Hamilton TJ, et al. Neuromorphic silicon neuron circuits. Front Neurosci 2011;5:1–23.
[46] Pickett MD, Medeiros-Ribeiro G, Williams RS. A scalable neuristor built with Mott memristors. Nat Mater 2013;12:114–7.
[47] Heck M, Bowers J. Energy efficient and energy proportional optical interconnects for multi-core processors: driving the need for on-chip sources. Select Top Quantum Electron IEEE J 2014;20:332–43.
[48] Liang D, Bowers JE. Recent progress in lasers on silicon. Nat Photon 2010;4:511–7.
[49] Roelkens G, Liu L, Liang D, et al. III-V/silicon photonics for on-chip and intra-chip optical interconnects. Laser Photon Rev 2010;4:751–79.
[50] Vlasov Y. Silicon CMOS-integrated nano-photonics for computer and data communications beyond 100g. Commun Mag IEEE 2012;50:s67–72.
[51] Barwicz T, Boyer N, Harel S, et al. Automated, self-aligned assembly of 12 fibers per nanophotonic chip with standard microelectronics assembly tooling. In: Electronic Components and Technology Conference (ECTC), 2015 IEEE 65th, 2015:775–82.
[52] Sysak M, Liang D, Jones R, et al. Hybrid silicon laser technology: a thermal perspective. Select Top Quantum Electron IEEE J 2011;17:1490–8.
[53] Yamada M. A theoretical analysis of self-sustained pulsation phenomena in narrow-stripe semiconductor lasers. IEEE J Quantum Electron 1993;29:1330–6.
[54] Dubbeldam JLA, Krauskopf B. Self-pulsations of lasers with saturable absorber: dynamics and bifurcations. Opt Commun 1999;159:325–38.
[55] Dubbeldam JLA, Krauskopf B, Lenstra D. Excitability and coherence resonance in lasers with saturable absorber. Phys Rev E 1999;60:6580–8.
[56] Elsass T, Gauthron K, Beaudoin G, Sagnes I, Kuszelewicz R, Barbay S. Control of cavity solitons and dynamical states in a monolithic vertical cavity laser with saturable absorber. Eur Phys J D 2010;59:91–6.
[57] Larotonda MA, Hnilo A, Mendez JM, Yacomotti AM. Experimental investigation on excitability in a laser with a saturable absorber. Phys Rev A 2002;65:033812.
[58] Nahmias MA, Tait AN, Shastri BJ, de Lima TF, Prucnal PR. Excitable laser processing network node in hybrid silicon: analysis and simulation. Opt Express 2015;23:26800–13.
[59] Shastri BJ, Nahmias MA, Tait AN, Prucnal PR. Simulations of a graphene excitable laser for spike processing. Opt Quantum Electron 2014;46:1353–8.

[60] Shastri BJ, Nahmias MA, Tait AN, Wu B, Prucnal PR. Simpel: circuit model for photonic spike processing laser neurons. Opt Express 2015;23:8029–44.
[61] Spühler GJ, Paschotta R, Fluck R, et al. Experimentally confirmed design guidelines for passively q-switched microchip lasers using semiconductor saturable absorbers. J Opt Soc Am B Opt Phys 1999;16:376–88.
[62] Coomans W, Beri S, Sande GVD, Gelens L, Danckaert J. Optical injection in semiconductor ring lasers. Phys Rev A 2010;81:033802.
[63] Coomans W, Gelens L, Beri S, Danckaert J, Van Der Sande G. Solitary and coupled semiconductor ring lasers as optical spiking neurons. Phys Rev E Stat Nonlinear Soft Matter Phys 2011;84:1–8.
[64] Coomans W, Van der Sande G, Gelens L. Oscillations and multistability in two semiconductor ring lasers coupled by a single waveguide. Phys Rev A 2013;88:033813.
[65] Gelens L, Mashal L, Beri S, et al. Excitability in semiconductor microring lasers: experimental and theoretical pulse characterization. Phys Rev A 2010;82:063841.
[66] Van Vaerenbergh T, Alexander K, Dambre J, Bienstman P. Excitation transfer between optically injected microdisk lasers. Opt Express 2013;21:28922.
[67] Brunstein M, Yacomotti AM, Sagnes I, Raineri F, Bigot L, Levenson A. Excitability and self-pulsing in a photonic crystal nanocavity. Phys Rev A 2012;85:031803.
[68] Yacomotti AM, Monnier P, Raineri F, et al. Fast thermo-optical excitability in a two-dimensional photonic crystal. Phys Rev Lett 2006;97:143904.
[69] Yacomotti AM, Raineri F, Vecchi G, et al. All-optical bistable band-edge Bloch modes in a two-dimensional photonic crystal. Appl Phys Lett 2006;88.
[70] Romeira B. Dynamics of resonant tunneling diode optoelectronic oscillators. PhD thesis, Universidade do Algarve, 2012.
[71] Romeira B, Avó R, Javaloyes J, Balle S, Ironside C, Figueiredo J. Stochastic induced dynamics in neuromorphic optoelectronic oscillators. Opt Quantum Electron 2014;46:1391–6.
[72] Romeira B, Javaloyes J, Ironside CN, Figueiredo JML, Balle S, Piro O. Excitability and optical pulse generation in semiconductor lasers driven by resonant tunneling diode photo-detectors. Opt Express 2013;21:20931–40.
[73] Barland S, Piro O, Giudici M, Tredicce JR, Balle S. Experimental evidence of van der Pol-Fitzhugh-Nagumo dynamics in semiconductor optical amplifiers. Phys Rev E 2003;68:036209.
[74] Garbin B, Goulding D, Hegarty SP, Huyet G, Kelleher B, Barland S. Incoherent optical triggering of excitable pulses in an injection-locked semiconductor laser. Opt Lett 2014;39:1254.
[75] Garbin B, Javaloyes J, Tissoni G, Barland S. Topological solitons as addressable phase bits in a driven laser. Nat Commun 2015;6:5915.
[76] Goulding D, Hegarty SP, Rasskazov O, et al. Excitability in a quantum dot semiconductor laser with optical injection. Phys Rev Lett 2007;98:153903.
[77] Kelleher B, Bonatto C, Huyet G, Hegarty SP. Excitability in optically injected semiconductor lasers: contrasting quantum-well- and quantum-dot-based devices. Phys Rev E Stat Nonlinear Soft Matter Phys 2011;83:1–6.
[78] Kelleher B, Bonatto C, Skoda P, Hegarty SP, Huyet G. Excitation regeneration in delay-coupled oscillators. Phys Rev E Stat Nonlinear Soft Matter Phys 2010;81:1–5.
[79] Marino F, Balle S. Excitable optical waves in semiconductor microcavities. Phys Rev Lett 2005;94:094101.
[80] Turconi M, Garbin B, Feyereisen M, Giudici M, Barland S. Control of excitable pulses in an injection-locked semiconductor laser. Phys Rev E 2013;88:022923.
[81] Wieczorek S, Krauskopf B, Lenstra D. Unifying view of bifurcations in a semiconductor laser subject to optical injection. Opt Commun 1999;172:279–95.
[82] Wieczorek S, Krauskopf B, Lenstra D. Multipulse excitability in a semiconductor laser with optical injection. Phys Rev Lett 2002;88:063901.
[83] Wieczorek S, Krauskopf B, Simpson TB, Lenstra D. The dynamical complexity of optically injected semiconductor lasers. Phys Rep 2005;416:1–128.
[84] Aragoneses A, Perrone S, Sorrentino T, Torrent MC, Masoller C. Unveiling the complex organization of recurrent patterns in spiking dynamical systems. Sci Rep 2014;4:4696 EP.
[85] Giacomelli G, Giudici M, Balle S, Tredicce JR. Experimental evidence of coherence resonance in an optical system. Phys Rev Lett 2000;84:3298–301.
[86] Heil T, Fischer I, Elsäßer W, Gavrielides A. Dynamics of semiconductor lasers subject to delayed optical feedback: the short cavity regime. Phys Rev Lett 2001;87:243901.
[87] Giudici M, Green C, Giacomelli G, Nespolo U, Tredicce JR. Andronov bifurcation and excitability in semiconductor lasers with optical feedback. Phys Rev E 1997;55:6414–8.
[88] Sorrentino T, Quintero-Quiroz C, Aragoneses A, Torrent MC, Masoller C. Effects of periodic forcing on the temporally correlated spikes of a semiconductor laser with feedback. Opt Express 2015;23:5571–81.
[89] Wünsche HJ, Brox O, Radziunas M, Henneberger F. Excitability of a semiconductor laser by a two-mode homoclinic bifurcation. Phys Rev Lett 2001;88:023901.
[90] Yacomotti AM, Eguia MC, Aliaga J, Martinez OE, Mindlin GB, Lipsich A. Interspike time distribution in noise driven excitable systems. Phys Rev Lett 1999;83:292–5.
[91] Hurtado A, Henning ID, Adams MJ. Optical neuron using polarisation switching in a 1550nm-VCSEL. Opt Express 2010;18:25170–6.
[92] Hurtado A, Schires K, Henning ID, Adams MJ. Investigation of vertical cavity surface emitting laser dynamics for neuromorphic photonic systems. Appl Phys Lett 2012;100:103703.
[93] Prucnal PR, Shastri BJ, Ferreira de Lima T, Nahmias MA, Tait AN. Recent progress in semiconductor excitable lasers for photonic spike processing. Adv Opt Photon 2016;8:228.
[94] Durstewitz D, Seamans JK, Sejnowski TJ. Neurocomputational models of working memory. Nat Neurosci 2000;3:1184–91.
[95] Pillow JW, Shlens J, Paninski L, et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature 2008;454:995–9.
[96] Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Netw Comput Neural Syst 2001;12:289–316.
[97] Izhikevich EM. Polychronization: computation with spikes. Neural Comput 2006;18:245–82.
[98] Chrostowski L, Hochberg M. Silicon photonics design: from devices to systems. Cambridge University Press, Cambridge, UK, 2015.

[99] Smit MK. Generic InP-based integration technology, today and tomorrow. In: Advanced Photonics Congress. Washington, DC, USA, 2012:IM2A.1.
[100] Tait AN, Wu AX, de Lima TF, et al. Microring weight banks. IEEE J Select Top Quantum Electron 2016;22:312–25.
[101] Wang Y, Wang X, Flueckiger J, et al. Focusing sub-wavelength grating couplers with low back reflections for rapid prototyping of silicon photonic circuits. Opt Express 2014;22:20652–62.
[102] Tait A, Ferreira de Lima T, Nahmias M, Shastri B, Prucnal P. Continuous calibration of microring weights for analog optical networks. Photon Technol Lett IEEE 2016;28:887–90.
[103] Preston K, Sherwood-Droz N, Levy JS, Lipson M. Performance guidelines for WDM interconnects based on silicon microring resonators. In: CLEO:2011 - Laser Applications to Photonic Applications. Optical Society of America, 2011:CThP4.
[104] Ramaswami R. Multiwavelength lightwave networks for computer communication. Commun Mag IEEE 1993;31:78–88.
[105] Klein E, Geuzebroek D, Kelderman H, Sengo G, Baker N, Driessen A. Reconfigurable optical add-drop multiplexer using microring resonators. Photon Technol Lett IEEE 2005;17:2358–60.
[106] Mak J, Sacher W, Xue T, Mikkelsen J, Yong Z, Poon J. Automatic resonance alignment of high-order microring filters. Quantum Electron IEEE J 2015;51:1–11.
[107] Cox JA, Lentine AL, Trotter DC, Starbuck AL. Control of integrated micro-resonator wavelength via balanced homodyne locking. Opt Express 2014;22:11279–89.
[108] Cardenas J, Foster MA, Sherwood-Droz N, Poitras CB, Lira HLR, Zhang B, et al. Wide-bandwidth continuously tunable optical delay line using silicon microring resonators. Opt Express 2010;18:26525–34.
[109] DeRose CT, Watts MR, Trotter DC, Luck DL, Nielson GN, Young RW. Silicon microring modulator with integrated heater and temperature sensor for thermal control. In: Conference on Lasers and Electro-Optics 2010. Optical Society of America, 2010:CThJ3.
[110] Jayatilleka H, Murray K, Ángel Guillén-Torres M, et al. Wavelength tuning and stabilization of microring-based filters using silicon in-resonator photoconductive heaters. Opt Express 2015;23:25084–97.
[111] Akopyan F, Sawada J, Cassidy A, et al. Truenorth: design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. Comput Aided Des Integr Circuits Syst IEEE Trans 2015;34:1537–57.
[112] Tait A, Nahmias M, Ferreira de Lima T, et al. Continuous control of microring weight banks. In: Proc. IEEE Photonics Conf. (IPC), 2015.
[113] Tait AN, Ferreira de Lima T, Nahmias MA, Shastri BJ, Prucnal PR. Multi-channel control for microring weight banks. Opt Express 2016;24:8895–906.
[114] Green500 list. https://www.top500.org/green500/. June 2016.
[115] Friedmann S, Frémaux N, Schemmel J, Gerstner W, Meier K. Reward-based learning under hardware constraints – using a RISC processor embedded in a neuromorphic substrate. Front Neurosci 2013;7:160.
[116] Jayatilleka H, Murray K, Caverley M, Jaeger N, Chrostowski L, Shekhar S. Crosstalk in SOI microring resonator-based filters. Lightw Technol J 2015;34:2886–96.
[117] Sherwood-Droz N, Preston K, Levy JS, Lipson M. Device guidelines for WDM interconnects using silicon microring resonators. In: Workshop on the Interaction between Nanophotonic Devices and Systems (WINDS), colocated with Micro. Vol. 43. 2010:15–8.
[118] Xu Q, Fattal D, Beausoleil RG. Silicon microring resonators with 1.5-μm radius. Opt Express 2008;16:4309–15.
[119] Biberman A, Shaw MJ, Timurdogan E, Wright JB, Watts MR. Ultralow-loss silicon ring resonators. Opt Lett 2012;37:4236–8.
[120] Xiong K, Xiao X, Hu Y, et al. Single-mode silicon-on-insulator elliptical microdisk resonators with high q factors. In: Photonics and Optoelectronics Meetings (POEM). 2011;8333:83330A–A-7.
[121] Soltani M, Li Q, Yegnanarayanan S, Adibi A. Toward ultimate miniaturization of high Q silicon traveling-wave microresonators. Opt Express 2010;18:19541–57.
[122] Ferreira de Lima T, Shastri BJ, Nahmias MA, Tait AN, Prucnal PR. Physical modeling of photonic neural networks. In: Summer Topicals Meeting Series (SUM), 2016. IEEE, 2016.
[123] Jalali B, Mahjoubfar A. Tailoring wideband signals with a photonic hardware accelerator. Proc IEEE 2015;103:1071–86.
[124] Brunner D, Soriano MC, Mirasso CR, Fischer I. Parallel photonic information processing at gigabyte per second data rates using transient states. Nat Commun 2013;4:1364.
[125] Duport F, Schneider B, Smerieri A, Haelterman M, Massar S. All-optical reservoir computing. Opt Express 2012;20:22783–95.
[126] Larger L, Soriano MC, Brunner D, et al. Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing. Opt Express 2012;20:3241–9.
[127] Ortn S, Soriano MC, Pesquera L, et al. A unified framework for reservoir computing and extreme learning machines based on a single time-delayed neuron. Sci Rep 2015;5:14945 EP.
[128] Vandoorne K, Mechet P, Van Vaerenbergh T, et al. Experimental demonstration of reservoir computing on a silicon photonics chip. Nat Commun 2014;5:3541.
[129] Larsson E, Edfors O, Tufvesson F, Marzetta T. Massive MIMO for next generation wireless systems. IEEE Commun Mag 2014;52:186–95.
[130] Gesbert D, Shafi M, Shan Shiu D, Smith PJ, Naguib A. From theory to practice: an overview of MIMO space-time coded wireless systems. IEEE J Select Areas Commun 2003;21:281–302.
[131] Hansen RC. Phased array antennas. Vol. 213. John Wiley & Sons, Hoboken, NJ, USA, 2009.
[132] Jerez JL, Constantinides GA, Kerrigan EC. An FPGA implementation of a sparse quadratic programming solver for constrained predictive control. In: ACM/SIGDA International Symposium on Field Programmable Gate Arrays – FPGA, 2011:209–18.
[133] Xia Y. A new neural network for solving linear and quadratic programming problems. IEEE Trans Neural Netw 2001;12:1074–83.
[134] Keviczky T, Balas GJ. Receding horizon control of an F-16 aircraft: a comparative study. Control Eng Pract 2006;14:1023–33.