Neuromorphic Photonics, Principles of

Bhavin J. Shastri, Alexander N. Tait, Thomas Ferreira de Lima, Mitchell A. Nahmias, Hsuan-Tung Peng and Paul R. Prucnal
Department of Electrical Engineering, Princeton University, Princeton, NJ, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Neuromorphic Computing: Beyond von Neumann and Moore
Technology Platforms
Neuromorphic Photonics
Photonic Neuron
Excitable/Spiking Lasers
Photonic Neural Network Architecture
Neuromorphic Platform Comparison
Future Directions
Summary and Conclusion
Bibliography

Glossary

Benchmark A standardized task that can be performed by disparate computing approaches, used to assess their relative processing merit in specific cases.

Bifurcation A qualitative change in behavior of a dynamical system in response to parameter variation. Examples include cusp (from monostable to bistable), Hopf (from stable to oscillating), and transcritical (exchange of stability between two steady states).

Brain-inspired computing (a.k.a. neuro-inspired computing) A biologically inspired approach to build processors, devices, and computing models for applications including adaptive control, machine learning, and cognitive radio. Similarities with biological signal processing include architectural, such as distributed; representational, such as analog or spiking; or algorithmic, such as adaptation.

Broadcast and Weight A multiwavelength analog networking protocol in which multiple all photonic neuron outputs are multiplexed and distributed to all neuron inputs. Weights are reconfigured by tunable spectral filters.

Excitability A far-from-equilibrium nonlinear dynamical mechanism underlying all-or-none responses to small perturbations.

Fan-in The number of inputs to a neuron.

Layered network A network topology consisting of a series of sets (i.e., layers) of neurons. The neurons in each set project their outputs only to neurons in the subsequent layer. Most commonly used type of network used for machine learning.

Metric A quantity assessing performance of a device in reference to a specific computing approach.

Microring weight bank A silicon photonic implementation of a reconfigurable spectral filter capable of independently setting transmission at multiple carrier wavelengths.

Modulation The act of representing an abstract variable in a physical quantity, such as photon rate (i.e., optical power), free carrier density (i.e., optical gain), and carrier drift (i.e., current). Electro-optic modulators are devices that convert from an electrical signal to the power envelope of an optical signal.

Moore's law An observation that the number of transistors in an integrated circuit doubles every 18 to 24 months, doubling its performance.

Multiply-accumulate (MAC) A common operation that represents a single multiplication followed by an addition: $a \leftarrow a + (b \times c)$.

Neural networks A wide class of computing models consisting of a distributed set of nodes, called neurons, interconnected with configurable or adaptable strengths, called weights. Overall neural network behavior can

R. A. Meyers (ed.), *Encyclopedia of Complexity and Systems Science*,
https://doi.org/10.1007/978-3-642-27737-5_702-1

be extremely complex relative to single neuron behavior.

Neuromorphic computing Computing approaches based on specialized hardware that formally adheres to one or more neural network models. Algorithms, metrics, and benchmarks can be shared between disparate neuromorphic computers that adhere to a common mathematical model.

Neuromorphic photonics An emerging field at the nexus of photonics and neural network processing models, which combines the complementary advantages of optics and electronics to build systems with high efficiency, high interconnectivity, and extremely high bandwidth.

Optoelectronics A technology of electronic devices and systems (semiconductor lasers, photodetectors, modulators, photonic integrated circuits) that interact (source, detect, and control).

Photonic integrated circuits (PICs) A chip that integrates many photonic components (lasers, modulators, filters, detectors) connected by optical waveguides that guide light; similar to an electronic integrated circuit that consists of transistors, diodes, resistors, capacitors, and inductors, connected by conductive wires.

Physical cascadability The ability of one neuron to produce an output with the same representational properties as its inputs. For example, photonic-electronicphotonic or 32bit-analog-32bit.

Recurrent network A network topology in which each neuron output can reach every other neuron, including itself. Every network is a subset of a recurrent network.

Reservoir computing A computational approach in which a complex, nonlinear substrate performs a diversity of functions, from which linear classifiers extract the most useful information to perform a given algorithm. The reservoir can be implemented by a recurrent neural network or a wide variety of other systems, such as time-delayed feedback.

Semiconductor lasers Lasers based on semiconductor gain media, where optical gain is achieved by stimulated emission at an interband transition under conditions of a high carrier density in the conduction band.

Signal cascadability The ability of one neuron to elicit an equivalent or greater response when driving multiple other neurons. The number of target neurons is called fan-out.

Silicon photonics A chip-scale, silicon-on-insulator (SOI) platform for monolithic integration of optics and microelectronics for guiding, modulating, amplifying, and detecting light.

Spiking neural networks (SNNs) A biologically realistic neural network model that processes information with spikes or pulses that encode information temporally.

Wavelength-division multiplexing (WDM) One of the most common multiplexing techniques used in optics where different wavelengths (colors) of light are combined, transmitted, and separated again.

WDM weighted addition A simultaneous summation of power modulated signals and transduction from multiple optical carriers to one electronic carrier. Occurs when multiplexed optical signals impinge on a standard photodetector.

Weight matrix A way to describe all network connection strengths between neurons arranged such that rows are input neuron indices and columns are output neuron indices. The weight matrix can be constrained to be symmetric, block off-diagonal, sparse, etc. to represent particular kinds of neural networks.

Weighted addition The operation describing how multiple inputs to a neuron are combined into one variable. Can be implemented in the digital domain by multiply-accumulate (MAC) operations or in the analog domain by various physical processes (e.g., current summing, total optical power detection).

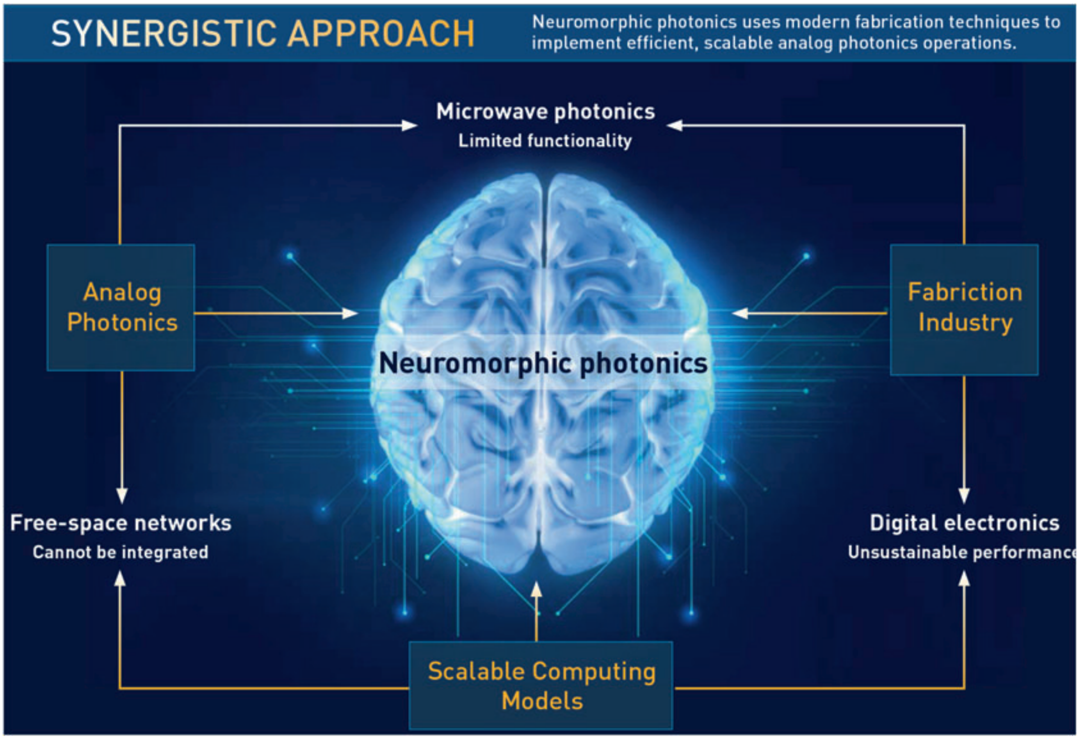
Definition of the Subject

In an age overrun with information, the ability to process reams of data has become crucial. The demand for data will continue to grow as smart gadgets multiply and become increasingly integrated into our daily lives. Next-generation industries in artificial intelligence services and high-performance computing are so far supported by microelectronic platforms. These data-intensive enterprises rely on continual improvements in hardware. Their prospects are running up against a stark reality: conventional one-size-fits-all solutions offered by digital electronics can no longer satisfy this need, as Moore's law (exponential hardware scaling), interconnection density, and the von Neumann architecture reach their limits.

With its superior speed and reconfigurability, analog photonics can provide some relief to these problems; however, complex applications of analog photonics have remained largely unexplored due to the absence of a robust photonic integration industry. 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.

Despite the advent of commercially viable photonic integration platforms, significant challenges still remain before scalable analog photonic processors can be realized. A central challenge is the development of mathematical bridges linking photonic device physics to models of complex analog information processing. Among such models, those of neural networks are perhaps the most widely studied and used by machine learning and engineering fields.

Recently, the scientific community has set out to build bridges between the domains of photonic device physics and neural networks, giving rise to the field of *neuromorphic photonics* (Fig. 1). This entry reviews the recent progress in integrated neuromorphic photonics. We provide an overview of neuromorphic computing, discuss the associated technology (microelectronic and photonic) platforms, and compare their metric performance.



Neuromorphic Photonics, Principles of, Fig. 1 The advent of neuromorphic photonics is due to the convergence of recent advances in photonic integration technology, resurgence of scalable computing models (e.g., spiking, deep neural networks), and a large-scale silicon industrial ecosystem

We discuss photonic neural network approaches and challenges for integrated neuromorphic photonic processors while providing an in-depth description of photonic neurons and a candidate interconnection architecture. We conclude with a future outlook of neuro-inspired photonic processing.

Introduction

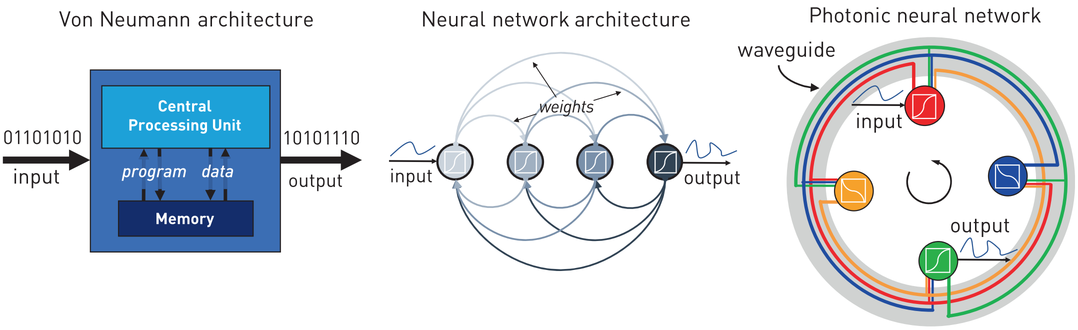
Complexity manifests in our world in countless ways (Strogatz 2001; Vicsek 2002) ranging from intracellular processes (Crescenzi et al. 1998) and human brain area function (Markram et al. 2011) to climate dynamics (Donges et al. 2009) and world economy (Hidalgo et al. 2007). An understanding of complex systems is a fundamental challenge facing the scientific community. Understanding complexity could impact the progress of our society as a whole, for instance, in fighting diseases, mitigating climate change, or creating economic benefits.

Current approaches to complex systems and big data analysis are based on software, executed on serialized and centralized von Neumann machines. However, the interconnected structure of complex systems (i.e., many elements interacting strongly and with variation (Bhalla and Iyengar 1999)) makes them challenging to reproduce in this conventional computing framework. Memory and data interaction bandwidths constrain the types of informatic systems that are feasible to simulate. The conventional computing approach has persisted thus far due to an exponential performance scaling in digital electronics, most famously embodied in Moore's law. For most of the past 60 years, the density of transistors, clock speed, and power efficiency in microprocessors have approximately doubled every 18 months. These empirical laws are fundamentally unsustainable. The last decade has witnessed a statistically significant (>99.95% likelihood (Marr et al. 2013)) plateau in processor energy efficiency.

This situation suggests that the time is ripe for radically new approaches to information processing, one being *neuromorphic photonics*. An emerging field at the nexus of photonics and neuroscience, neuromorphic photonics combines the complementary advantages of optics and electronics to build systems with high efficiency, high interconnectivity, and extremely high bandwidth. This entry reviews the recent progress in integrated neuromorphic photonic research. First, we introduce neuromorphic computing and give an overview of current technology platforms in microelectronics and photonics. Next, we discuss the evolution of neuromorphic photonics, the present photonic neural network approaches, and challenges for emerging integrated neuromorphic photonic processors. We introduce the concept of a "photonic neuron" followed by a discussion on its feasibility. We summarize recent research on optical devices that could be used as network-compatible photonic neurons. We discuss a networking architecture that efficiently channelizes the transmission window of an integrated waveguide. We also compare metrics between neuromorphic electronics and neuromorphic photonics. Finally, we offer a glimpse at the future outlook for neuro-inspired photonic processing and discuss potential applications.

Neuromorphic Computing: Beyond von Neumann and Moore

Conventional digital computers are based on the von Neumann architecture (von Neumann 1993). It consists of a memory that stores both data and instructions, a central processing unit (CPU), and inputs and outputs (Fig. 2 (left)). Instructions and data stored in the memory unit are separated from the CPU by a shared digital bus. This is known as the von Neumann bottleneck (Backus 1978) which fundamentally limits the performance of the system – a problem that is aggravated as CPUs become faster and memory units larger. This computing paradigm has dominated for over 60 years driven in part by the continual progress dictated by Moore's law (Moore 2000) for CPU scaling and Koomey's law (Koomey et al. 2011) for energy efficiency (in multiply-accumulate (MAC) operations per joule) compensating the bottleneck. Over the last several years,



Neuromorphic Photonics, Principles of, Fig. 2 Neural nets: The photonic edge. Von Neumann architectures (left), relying on sequential input-output through a central processor, differ fundamentally from more decentralized neural network architectures (middle). Photonic neural nets (right) can solve the interconnect bottleneck by using one waveguide to carry signals from many connections (easily $N_2 \sim 10,000$) simultaneously

such scaling has not followed suit, approaching an asymptote. The computation efficiency levels off below 10 GMAC/s/W or 100 pJ per MAC (Hasler and Marr 2013). The reasons behind this trend can be traced to both the representation of information at the physical level and the interaction of processing with memory at the architectural level (Marr et al. 2013).

Breaching the energy efficiency wall of digital computation by orders of magnitude is not fundamentally impossible. In fact, the human brain, believed to be the most complex system in the universe, is estimated to execute an amazing 10^{18} MAC/s using only 20 W of power (Hasler and Marr 2013; Merkle 1989). It does this with 10^{11} neurons with an average of 10^4 inputs each. This leads to an estimated total of 10^{15} *synaptic* connections, all conveying signals up to 1 kHz bandwidth. The calculated computational efficiency for the brain (< aJ/MAC) is therefore eight orders of magnitude beyond that of current supercomputers (100pJ/MAC). The brain is a natural standard for information processing, one that has been compared to artificial processing systems since their earliest inception. Nevertheless, the brain as a processor differs radically from computers today, both at the physical level and at the architectural level. Its exceptional performance is, at least partly, due to the neuron biochemistry, its underlying architecture, and the biophysics of neuronal computation algorithms.

Neuromorphic computing offers hope to building large-scale "bio-inspired" hardware whose computational efficiencies move toward those of a human brain. In doing so, neuromorphic platforms (Fig. 2 (middle)) could break performance limitations inherent in traditional von Neumann architectures in solving particular classes of problems. Their distributed hardware architectures can most efficiently evaluate models with high data interconnection, among which are real-time complex system assurance and big data awareness.

At the device level, digital CMOS is reaching physical limits (Mathur 2002; Taur et al. 1997). As the CMOS feature sizes scale down below 90 nm to 65 nm, the voltage, capacitance, and delay no longer scale according to a well-defined rate by Dennard's law (Dennard et al. 1974). This leads to a trade-off between performance (when transistor is on) and subthreshold leakage (when it is off). For example, as the gate oxide (which serves as an insulator between the gate and channel) is made as thin as possible (1.2 nm, around five atoms thick Si) to increase the channel conductivity, a quantum mechanical phenomenon of electron tunneling (Taur 2002; Lee and Hu 2001) occurs between the gate and channel leading to increased power

consumption. The recent shift to multi-core 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. Fixed power budgets have necessitated so-called *dark silicon* strategies (Esmaeilzadeh et al. 2012). Projections for the 8 nm node indicate that over

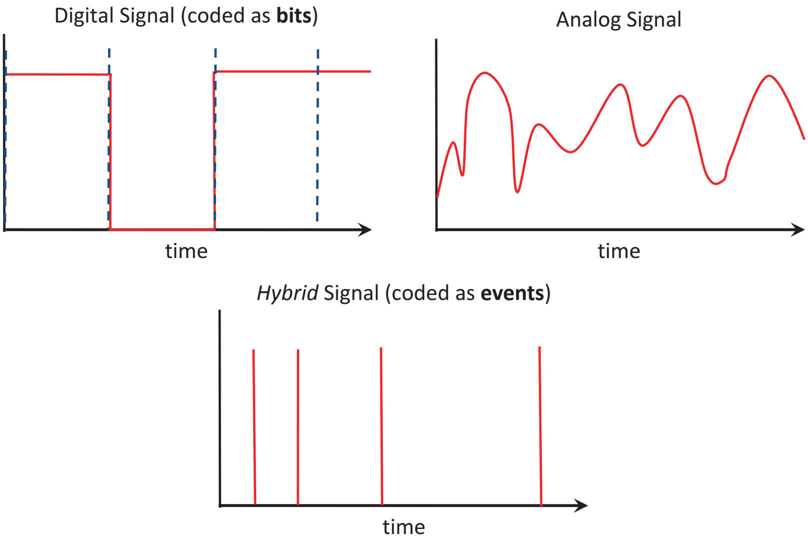
50% of the chip will be *dark* (Esmaeilzadeh et al. 2012), meaning unused at a given time. This has led to a widening rift between conventional computing capabilities and contemporary computing needs, particularly for the analysis of complex systems.

Computational tools have been revolutionary in hypothesis testing and simulation. They have led to the discovery of innumerable theories in science, and they will be an indispensable aspect of a holistic approach to problems in big data and many body physics; however, huge gaps between information structures in observed systems and standard computing architectures motivate a need for alternative paradigms if computational abilities are to be brought to the growing class of problems associated with complex systems. Brain-inspired computing approaches share the interconnected causal structures and dynamics analogous to the complex systems present in our world. This is in contrast to conventional approaches, where these structures are virtualized at considerable detriment to energy efficiency. From a usual constraint of a fixed power budget, energy efficiency caps overall simulation scale.

Over the years, there has been a deeply committed exploration of unconventional computing techniques (Hasler and Marr 2013; Keyes 1985; Jaeger and Haas 2004; Merolla et al. 2014; Modha et al. 2011; Tucker 2010; Caulfield and Dolev 2010; Woods and Naughton 2012; Benjamin et al. 2014; Pfeil et al. 2013; Furber et al. 2014; Snider 2007; Eliasmith et al. 2012; Indiveri et al. 2011; Brunner et al. 2013a; Tait et al. 2017; Shen et al. 2017; Shainline et al. 2017; Prucnal et al. 2016; Prucnal and Shastri 2017) to alleviate the device level and system/architectural level challenges faced by conventional computing platforms. Specifically, neuromorphic computing is going through an exciting period as it promises to make processors that use low energies while integrating massive amounts of information. Neuromorphic engineering aims to build machines employing basic nervous systems operations by bridging the physics of biology with engineering platforms enhancing performance for applications interacting with natural environments such as vision and speech (Hasler and Marr 2013). These neural-inspired systems are exemplified by a set of computational principles, including hybrid analog-digital signal representations (discussed next), co-location of memory and processing, unsupervised statistical learning, and distributed representations of information.

Technology Platforms

Information representation can have a profound effect on information processing. The *spiking* model found in biology is a sparse coding scheme recognized by the neuroscience community as a neural encoding strategy for information processing (Ostojic 2014; Paugam-Moisy and Bohte 2012; Kumar et al. 2010; Izhikevich 2003; Diesmann et al. 1999; Borst and Theunissen 1999) and has code-theoretic justifications (Sarpeshkar 1998; Thorpe et al. 2001; Maass et al. 2002). Spiking approaches promise extreme improvements in computational power efficiency (Hasler and Marr 2013) because they directly exploit the underlying physics of biology (Jaeger and Haas 2004; Sarpeshkar 1998; Maass 1997; Izhikivich 2007), analog electronics, or, in the present case, optoelectronics. Digital in amplitude but temporally analog, spiking naturally interleaves robust, discrete representations for communication with precise, continuous representations for computation in order to reap the benefits of both digital and analog. Spiking has two primary advantages over synchronous analog: (1) its analog variable (time) is less noisy than its digital variable (amplitude), and (2) it is asynchronous, without a global clock. Clock synchronization allows for time-division multiplexing (TDM); however, it is a significant practical problem in many-core systems (Mundy et al. 2015). These advantages may account for the ubiquity of spiking in natural processing systems (Thorpe et al. 2001). It is natural to deepen this distinction to include physical representational aspects, with the important result that optical carrier noise does not accumulate. As will be discussed, when an optical pulse is generated, it is transmitted and routed through a linear optical network with the help of its wavelength identifier (Fig. 3).

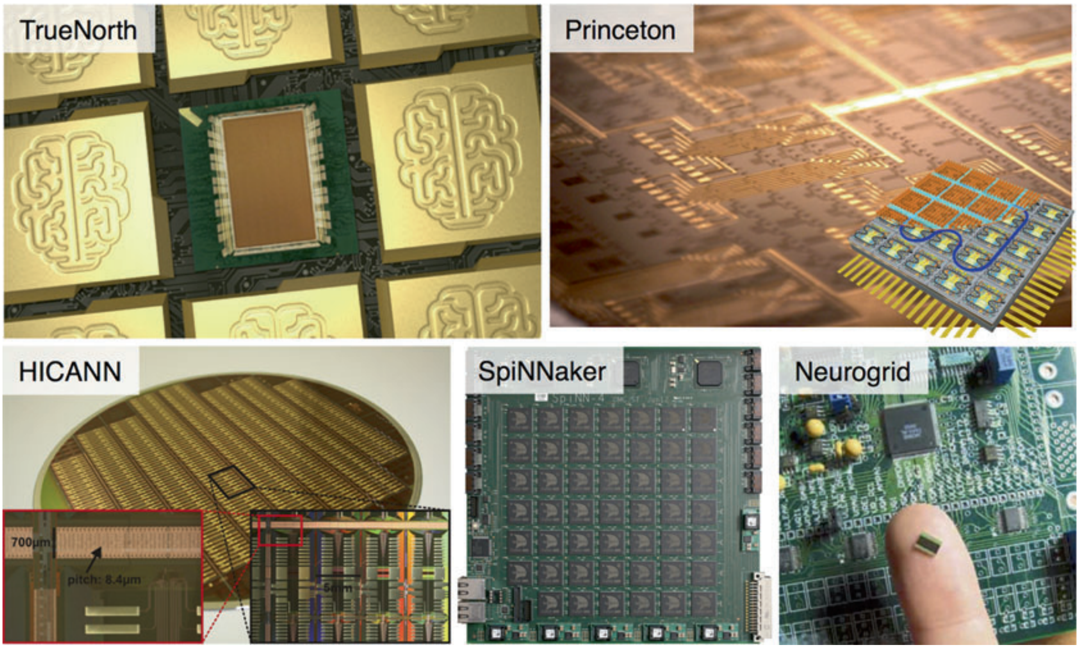

Digital Signal (coded as **bits**)
time
Analog Signal
time
Hybrid Signal (coded as **events**)
time

Neuromorphic Photonics, Principles of, Fig. 3 Spiking neural networks encode information as events in time rather than bits. Because the time at which a spike occurs is analog while its amplitude is digital, the signals use a mixed-signal or hybrid encoding scheme (Reproduced from Tait et al. 2014a). With permission of Springer)

Neuromorphic Microelectronics

Spiking primitives have been built in CMOS analog circuits, digital *neurosynaptic cores*, and non-CMOS devices. Various technologies (Fig. 4) have demonstrated large-scale spiking neural networks in electronics, including notably Neurogrid as part of Stanford University's Brains in Silicon program (Benjamin et al. 2014), IBM's TrueNorth as part of DARPA's SyNAPSE program (Merolla et al. 2014), HICANN as part of Heidelberg University's FACETS/BrainScaleS project (Schemmel et al. 2010), and University of Manchester's neuromorphic chip as part of the SpiNNaker project (Furber et al. 2014); the latter two are under the flagship of the European Commission's Human Brain Project (The HBP Report 2012). These spiking platforms promise potent advantages in efficiency, fault tolerance, and adaptability over von Neumann architectures to better interact with natural environments by applying the circuit and system principles of neuronal computation, including robust analog signaling, physics-based dynamics, distributed complexity, and learning. Using this neuromorphic hardware to process faster signals (e.g., radio waveforms) is, however, not a simple matter of accelerating the clock. These systems rely on slow timescale operation to accomplish dense interconnection.

Whereas von Neumann processors rely on point-to-point memory processor communication, a neuromorphic processor typically requires a large number of interconnects (i.e., ~100 s of many-to-one fan-in per processor) (Hasler and Marr 2013). This requires a significant amount of multicasting, which creates a communication burden. This, in turn, introduces fundamental performance challenges that result from RC and radiative physics in electronic links, in addition to the typical bandwidth-distance-energy limits of point-to-point connections (Miller 2000). While some incorporate a dense mesh of wires overlaying the semiconductor substrate as crossbar arrays, large-scale systems are ultimately forced to adopt some form of TDM or packet switching, notably, address-event representation (AER), which introduces the overhead of representing spike as digital codes instead of physical pulses. This abstraction at the architectural level allows virtual interconnectivities to exceed wire density by a factor related to the sacrificed bandwidth, which can be orders of magnitude (Boahen 2000). Spiking neural networks (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 (Merolla et al. 2014; Furber et al. 2014) and accelerated simulation in the low MHz regime (Schemmel et al. 2010). However,



Neuromorphic Photonics, Principles of, Fig. 4 Selected pictures of five different neuromorphic hardware discussed here. They include TrueNorth (Merolla et al. 2014), HICANN (Schemmel et al. 2010), SpiNNaker (Furber et al. 2014), Neurogrid (Benjamin et al. 2014)

neuromorphic processing for high-bandwidth applications in the GHz regime (such as sensing and manipulating the radio spectrum and for hypersonic aircraft control) must take a fundamentally different approach to interconnection.

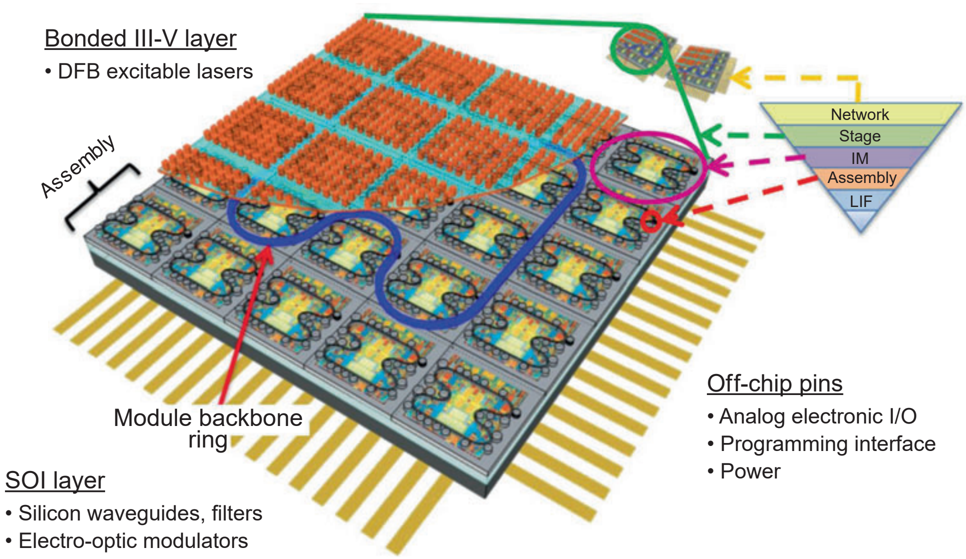
Toward Neuromorphic Photonics

Just as optics and photonics are being employed for interconnection in conventional CPU systems, optical networking principles can be applied to the neuromorphic domain (Fig. 2 (right)). Mapping a processing paradigm to its underlying dynamics, than abstracting the physics away entirely, can significantly streamline efficiency and performance, and mapping a laser's behavior to a neuron's behavior relies on discovering formal mathematical analogies (i.e., isomorphisms) in their respective governing dynamics. Many of the physical processes underlying photonic devices have been shown to have a strong analogy with biological processing models, which can both be described within the framework of nonlinear dynamics. Large-scale integrated photonic platforms (see Fig. 5) offer an opportunity for ultrafast neuromorphic processing that complements neuromorphic microelectronics aimed at biological timescales. The high switching speeds, high communication bandwidth, and low cross talk achievable in photonics are very well-suited for an ultrafast spike-based information scheme with high interconnection densities (Prucnal et al. 2016; Tait et al. 2014b). The efforts in this budding research field aims to synergistically integrate the underlying physics of photonics with spiking neuron-based processing. *Neuromorphic photonics* represents a broad domain of applications where quick, temporally precise, and robust systems are necessary.

Later in this entry, we compare metrics between neuromorphic electronics and neuromorphic photonics.

Neuromorphic Photonics

Photonics has revolutionized information transmission (communication and interconnects), while electronics, in parallel, has dominated



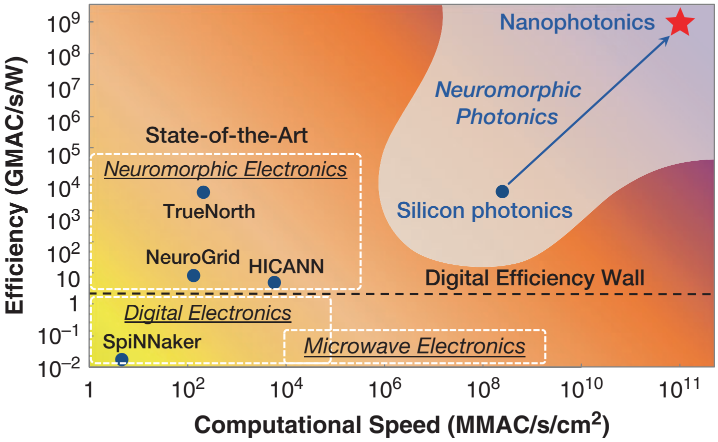
Neuromorphic Photonics, Principles of, Fig. 5 Conceptual rendering of a photonic neuromorphic processor. Laser arrays (orange layer) implement pulsed (spiking) dynamics with electro-optic physics, and a photonic network on-chip (blue and gray) supports complex structures of virtual interconnection among these elements, while electronic circuitry (yellow) controls stability, self-healing, and learning

information transformation (computations). This leads naturally to the following question: how can the unifying of the boundaries between the two be made as effective as possible? (Keyes 1985; Tucker 2010; Caulfield and Dolev 2010). CMOS gates only draw energy from the rail when and where called upon; however, the energy required to drive 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 (Prucnal and Shastri 2017).

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 (e.g., lasers and detectors) and passive elements (e.g., waveguides, resonators, modulators) is emerging (Fang et al. 2007; Liang et al. 2010; Liang and Bowers 2010a; Roelkens et al. 2010; Heck et al. 2013). A neuromorphic photonic approach based on this platform could potentially operate six to eight orders of magnitude faster than neuromorphic electronics when accounting for the bandwidth reduction of virtualizing interconnects (Prucnal and Shastri 2017) (cf. Fig. 6; also see Table 1 and related discussion).

In the past, the communication potentials of optical interconnects have received attention for neural networking; however, attempts to realize holographic or matrix-vector multiplication systems have failed to outperform mainstream electronics at relevant problems in computing, which can perhaps be attributed to the immaturity of large-scale integration technologies and manufacturing economics.



Neuromorphic Photonics, Principles of, Fig. 6 Comparison of neuromorphic hardware platforms. Neuromorphic photonic architectures potentially sport better speed-to-efficiency characteristics than state-of-the-art electronic neural nets (such as IBM's TrueNorth, Stanford University's Neurogrid, Heidelberg University's HICANN), as well as advanced digital electronic systems (such as the University of Manchester's SpiNNaker). On the top right: the photonic neuron platforms studied in Prucnal and Shastri (2017). The regions highlighted in the graph are approximate, based on qualitative trade-offs of each technology (Adapted from Ferreira de Lima et al. 2017). Licensed under Creative Commons Attribution-NonCommercial-NoDerivatives License (CC BY-NC-ND))

Neuromorphic Photonics, Principles of, Table 1 Comparison between different neuromorphic processors

Chip	MAC rate/ processor[a]	Energy/MAC (pJ)[b]	Processor fan-in	Area/MAC (μm^2)[c]	Synapse precision (bit)
Photonic hybrid III-V/Si[d]	20 GHz	0.26	108	205	5.1
Sub-λ photonics [e] (future trend)	200 GHz	0.0007	~200	20	8
HICANN (Schemmel et al. 2010)	22.4 MHz	198.4	224	780	4
TrueNorth (Merolla et al. 2014)	2.5 kHz	0.27	256	4.9	5
Neurogrid (Benjamin et al. 2014)	40.1 kHz	119	4096	7.1	13
SpiNNaker[f] (Furber et al. 2014)	3.2 kHz	6 × 10^5	320	217	16

[a] A 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

[b] The energy per MAC for HICANN, TrueNorth, Neurogrid, and SpiNNaker were estimated by dividing wall-plug power to number of neurons and to operational MAC rate per processor

[c] 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

[d] III–V/Si hybrid stands for estimated metrics of a spiking neural network in a photonic integrated circuit in a III–V/Si hybrid platform

[e] Sub-λ stands for estimated metrics for a platform using optimized sub-wavelength structures, such as photonic crystals

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

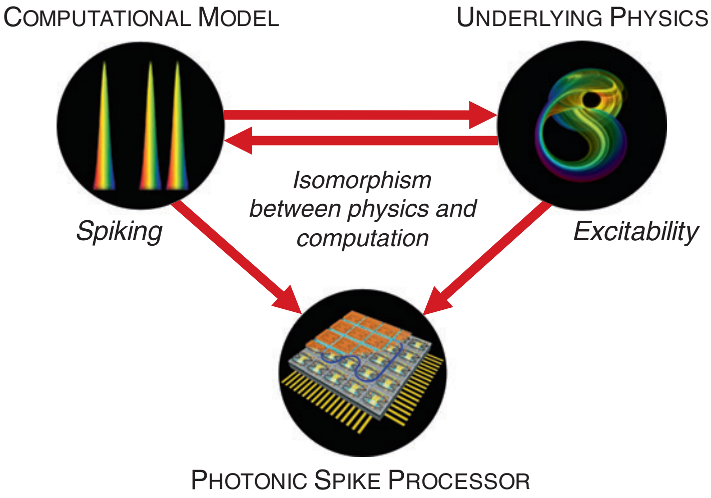
Techniques in silicon photonic integrated circuit (PIC) fabrication are driven by a tremendous demand for optical interconnects within conventional digital computing systems (Smit et al. 2012; Jalali and Fathpour 2006), which means platforms for systems integration of active photonics are becoming commercial reality (Liang et al. 2010; Roelkens et al. 2010; Heck et al. 2013; Liang and Bowers 2010b; Marpaung et al. 2013). The potential of recent advances in integrated photonics to enable unconventional computing has not yet been investigated. The theme of current research has been on how modern PIC platforms can topple historic technological barriers between large-scale analog networks and photonic neural systems. In this context, there are two complementary areas of investigation by the research community, namely, photonic spike processing and photonic reservoir computing. While the scope of this entry is limited to the former, we briefly introduce both.

Photonic Spike Processing

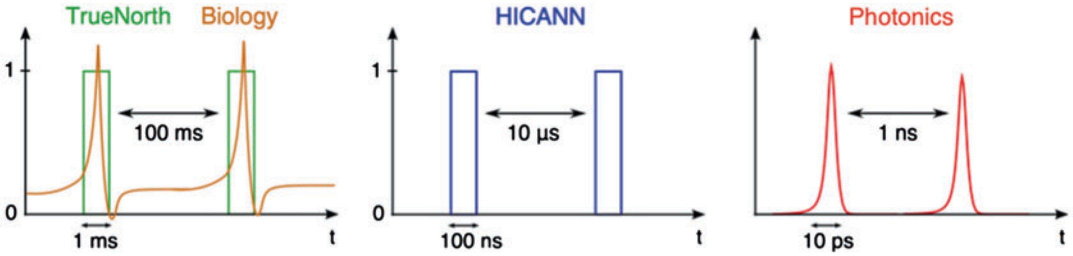
An investigation of photonics for information processing based on spikes has taken place alongside the development of electronic spiking architectures. Since the first demonstration of photonic spike processing by Rosenbluth et al. (Rosenbluth et al. 2009), there has been a surge in research related to aspects of spike processing in various photonic devices with a recent bloom of proposed forms of spiking dynamics (Tait et al. 2014b; Kelleher et al. 2010; Kravtsov et al. 2011; Fok et al. 2011; Coomans et al. 2011; Brunstein et al. 2012; Nahmias et al. 2013; Van Vaerenbergh et al. 2013; Aragoneses et al. 2014; Selmi et al. 2014; Hurtado and Javaloyes 2015; Garbin et al. 2015; Shastri et al. 2016; Romeira et al. 2016) – a strategy that could lead to combined computation and communication in the same substrate.

We recently (Prucnal et al. 2016; Prucnal and Shastri 2017) reviewed the recent surge of interest (Tait et al. 2014a; Kelleher et al. 2010; Coomans et al. 2011; Brunstein et al. 2012; Nahmias et al. 2013, 2015; Van Vaerenbergh et al. 2012, 2013; Aragoneses et al. 2014; Selmi et al. 2014; Hurtado and Javaloyes 2015; Garbin et al. 2015; Shastri et al. 2016; Yacomotti et al. 2006a; Goulding et al. 2007; Hurtado et al. 2010, 2012; Coomans 2012; Romeira et al. 2013, 2016; Sorrentino et al. 2015) in the information processing abilities of semiconductor devices that exploit the dynamical isomorphism between semiconductor photocarriers and neuron biophysics. Many of these proposals for "photonic neurons" or "laser neurons" or "optical neurons" for spike processing are based on lasers operating in an *excitable* regime (Fig. 7). Excitability (Hodgkin and Huxley 1952; Krauskopf et al. 2003) is a dynamical system property underlying all-or-none responses.

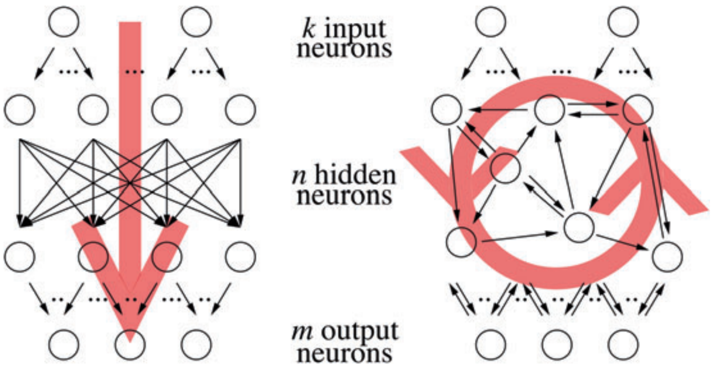
The difference in physical timescales allows these laser systems to exhibit these properties, except many orders of magnitude faster than



Neuromorphic Photonics, Principles of, Fig. 7 Analogies between spike processing and photonics can be exploited to create a computational paradigm that performs beyond the sum of its parts. By reducing the abstraction between process (spiking) and physics (excitability), there could be a significant advantage on speed, energy usage, scalability (Adapted with permission from Prucnal et al. 2016). Copyright 2016 Optical Society of America)



Neuromorphic Photonics, Principles of, Fig. 8 Difference in spike processing timescales (pulse width and refractory period) between biological neurons (left), electronic spiking neurons (middle), and photonic neurons (right) (Reproduced with permission from Prucnal et al. (2016). Copyright 2016 Optical Society of America)



Neuromorphic Photonics, Principles of, Fig. 9 Comparison of the architectures of a feedforward (left-hand side) with a recurrent neural network (right-hand side); the gray arrows sketch the possible direction of computation (Adapted from Burgsteiner (2005))

their biological counterparts (Nahmias et al. 2013); the temporal resolution (tied to spike widths) and processing speed (tied to refractory period) are accelerated by factors nearing 100 million (Fig. 8). A network of photonic neurons could open computational domains that demand unprecedented temporal precision, power efficiency, and functional complexity, potentially including applications in wideband radio-frequency (RF) processing, adaptive control of multi-antenna systems, and high-performance scientific computing.

Photonic Reservoir Computing

Reservoir computing (RC) is another promising approach for neuro-inspired computing. A central tenet of RC is that complex processes are generated in a medium whose behavior is not necessarily understood theoretically. Instead, the "reservoir" (a fixed, recurrent network of nonlinear nodes; see Fig. 9) generates a large number of complex processes, and linear combinations of reservoir signals are trained to approximate a desired task (Verstraeten et al. 2007). To arrive at a user-defined behavior, reservoirs do not need to model or program and instead rely on supervised machine learning techniques for simple linear classifiers. This is advantageous in systems whose overall behavior is complex, yet difficult to model or correspond to a theoretical behavior. There are a wide class of physical systems that fit that description, and the reservoir concept makes them highly likely to apply to widespread information processing tasks (Soriano et al. 2015).

Over the past several years, reservoir computers have been constructed that exploit the incredible bandwidths and speeds available to photonic signals. These "photonic reservoirs" utilize optical multiplexing strategies to form highly complex virtual networks. Photonic RC particularly is attractive when the information to be processed is already in the optical domain, for example, applications in telecommunications and image processing. Recently, there has been a significant development in the hardware realization of RC. Optical reservoirs have been demonstrated with various schemes such as benchtop demonstrations with a fiber with a single nonlinear dynamical node (Brunner et al. 2013a, b; Paquot et al. 2012; Martinenghi et al. 2012; Larger et al. 2012; Duport et al. 2012, 2016; Hicke et al. 2013; Soriano et al. 2013; Ortín et al. 2015), and integrated solutions including microring resonators (Mesaritakis et al. 2013), a network of coupled semiconductor optical amplifiers (SOAs) (Vandoorne et al. 2011), and a passive silicon photonic chip (Vandoorne et al. 2014).

It has been experimentally demonstrated and verified that some of these photonic RC solutions achieve highly competitive figures of merit at unprecedented data rates often outperforming software-based machine learning techniques for computationally hard tasks such as spoken digit and speaker recognition, chaotic time series prediction, signal classification, or dynamical system modeling. Another significant advantage of photonic-based approaches, as pointed out by Vandoorne et al. (Vandoorne et al. 2014), is the straightforward use of coherent light to exploit both the phase and amplitude of light. The simultaneous exploitation of two physical quantities results in a notable improvement over real-valued networks that are traditionally used in software-based RC – a reservoir operating on complex numbers in essence doubles the internal degrees of freedom in the system, leading to a reservoir size that is roughly twice as large as the same device operated with incoherent light.

Neuromorphic spike processing and reservoir approaches differ fundamentally and possess complementary advantages. Both derive a broad repertoire of behaviors (often referred to as complexity) from a large number of physical degrees of freedom (e.g., intensities) coupled through interaction parameters (e.g., transmissions). Both offer means of selecting a specific, desired behavior from this repertoire using controllable parameters. In neuromorphic systems, network weights are *both* the interaction and controllable parameters, whereas, in reservoir computers, these two groups of parameters are separate. This distinction has two major implications. Firstly, the interaction parameters of a reservoir do not need to be observable or even repeatable from system to system. Reservoirs can thus derive complexity from physical processes that are difficult to model or reproduce. Furthermore, they do not require significant hardware to control the state of the reservoir. Neuromorphic hardware has a burden to correspond physical parameters (e.g., drive voltages) to model parameters (e.g., weights). Secondly, reservoir computers can only be made to elicit a desired behavior through instance-specific supervised training, whereas neuromorphic computers can be programmed a priori using a known set of weights. Because neuromorphic behavior is determined only by controllable parameters, these parameters can be mapped directly between different system instances, different types of neuromorphic systems, and simulations. Neuromorphic hardware can leverage existing algorithms (e.g., Neural Engineering Framework (NEF) (Stewart and Eliasmith 2014)), map virtual training results to hardware, and particular behaviors are guaranteed to occur. Photonic RCs can of course be simulated; however, they have no corresponding guarantee that a particular hardware instance will reproduce a simulated behavior or that training will be able to converge to this behavior.

Challenges for Integrated Neuromorphic Photonics

Key criteria for nonlinear elements to enable a scalable computing platform include thresholding, fan-in, and cascadability (Keyes 1985; Tucker 2010; Caulfield and Dolev 2010). Past approaches to optical computing have met challenges realizing these criteria. A variety of digital logic gates in photonics that suppress amplitude noise accumulation have been claimed, but many proposed optical logic devices do not meet necessary conditions of cascadability. Analog photonic processing has found application in high-bandwidth filtering of microwave signals (Capmany et al. 2006), but the accumulation of phase noise, in addition to amplitude noise, limits the ultimate size and complexity of such systems.

Recent investigations (Prucnal and Shastri 2017; Tait et al. 2014b; Coomans et al. 2011; Brunstein et al. 2012; Nahmias et al. 2013, 2015; Van Vaerenbergh et al. 2013; Aragoneses et al.

2014; Selmi et al. 2014; Hurtado and Javaloyes 2015; Garbin et al. 2015; Shastri et al. 2016; Romeira et al. 2016) have suggested that an alternative approach to exploit the high bandwidth of photonic devices for computing lies not in increasing device performance or fabrication, but instead in examining models of computation that hybridize techniques from analog and digital processing. These investigations have concluded that a photonic neuromorphic processor could satisfy them by implementing a model of a neuron, i.e., a photonic neuron, as opposed to the model of a logic gate. Early work in neuromorphic photonics involved fiber-based spiking approaches for learning, pattern recognition, and feedback (Rosenbluth et al. 2009; Kravtsov et al. 2011; Fok et al. 2011). Spiking behavior resulted from a combination of 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. There are two research challenges necessary to establish this isomorphism: the nonlinearity (equivalent to thresholding) in individual neurons and the synaptic interconnection (related to fan-in and cascadability) between different neurons, as will be discussed in the proceeding sections. 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.

Photonic neurons 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) and 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.

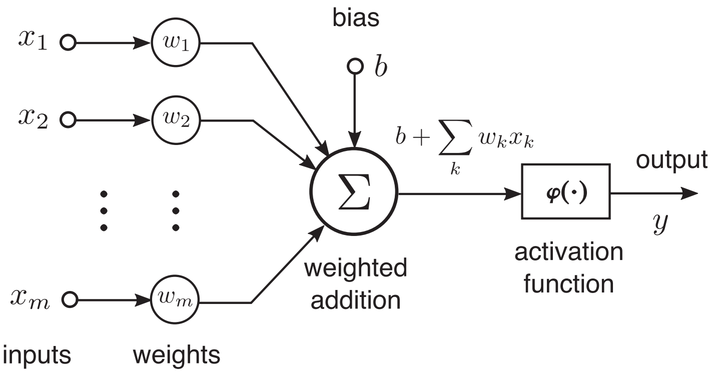
The optical channel is highly expressive and correspondingly very sensitive to phase and frequency noise. For example, the networking architecture proposed in Tait et al. (2014b) relies on wavelength-division multiplexing (WDM) for interconnecting many points in a photonic substrate together. Any proposal for networking computational primitives 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 addition, as will be discussed next, schemes which exhibit limitations with regard to wavelength channels may require a large number of wavelength conversion steps, which can be costly, noisy, and inefficient.

Photonic Neuron

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 (Bengio et al. 2013).

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 Fig. 10. 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



Neuromorphic Photonics, Principles of, Fig. 10 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 (Reproduced with permission from Prucnal et al. (2016). Copyright 2016 Optical Society of America)

addition) of the outputs of the neurons connected to it. Then, the particular neuron integrates the combined signal and produces a nonlinear response, represented by an *activation function*, usually monotonic and bounded.

Three generations of neural networks were historically studied in computational neuroscience (Maass 1997). The first was based on the McCulloch-Pitts neural model, which consists of a linear combiner followed by a steplike activation function (binary output). These neural networks are Boolean complete – i.e., 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 they can uniformly approximate arbitrarily well any continuous function with a compact domain (Maass 1997). When augmented with the notion of "time," recurrent connections can be created and be exploited to create attractor states (Eliasmith 2005) and associative memory (Hopfield 1982) 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 monkey 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 three spikes in 30 ms (Thorpe et al. 2001; Maass 1997; Perrett et al. 1982) 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 et al. 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 (Maass 1997).

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

$$C_m \frac{\mathrm{d}V_m(t)}{\mathrm{d}t} = \frac{1}{R_m}(V_m(t) - V_L) + I_{app}(t);$$
$$\text{if } V_m(t)\times > V_{\text{thresh}}$$
$$\text{then release a spike and set } V_m(t)\times \to V_{\text{reset}}, \tag{1}$$

where $V_m(t)$ is the membrane voltage, R_m the membrane resistance, V_L the equilibrium potential, and I_{app} the applied current (input). More bio-realistic 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. Bio-realistic SNN present a particular challenge because of the need for fine-grained time discretization (Izhikevich 2004). Engineers circumvent this challenge by employing an event-driven simulation model which 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.

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

Here, we 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.

Basic Requirements for a Photonic Neuron

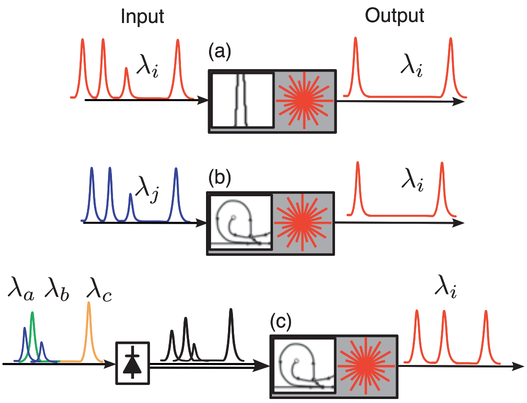
An artificial neuron described in Fig. 10 must perform three basic mathematical operations: array multiplication (weighting), summation, and a nonlinear transformation (activation function). Moreover, the inputs to be weighted in the first stage must be of the same nature of the output – in the case 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 higher 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.

The Processing Network Node

A networkable photonic device with optical I/O, provided that it is capable of emulating an artificial neuron, is named a processing network node (PNN) (Tait et al. 2014b). 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.

Silicon waveguides have a relatively enormous transparency window of 7.5 THz (Agrawal 2002) over which they can guide lightwaves with very



Neuromorphic Photonics, Principles of, Fig. 11 General classification of semiconductor excitable lasers based on (**a**) coherent optical injection, (**b**) non-coherent optical injection, and (**c**) full electrical injection. Each of these lasers can be pumped either electrically or optically (Reproduced from Ferreira de Lima et al. (2017). Licensed under Creative Commons Attribution-NonCommercial-NoDerivatives License. (CC BY-NC-ND))

low attenuation and cross talk, in contrast with electrical wires or radio-frequency transmission lines. With WDM, each input signal exists at a different wavelength but is superimposed with other signals onto the same waveguide. For example, to maximize the information throughput, a single waveguide could carry hundreds of wide-band signals (~20 GHz) simultaneously. As such, 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 other hand, 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 the next section.

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 (Fig. 11a). Since coherent optical systems operate at a single wavelength λ, the signals lack distinguishability from one another in a WDM-encoded framework. As demonstrated in Alexander et al. (2013), 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 (Fig. 11b). A number of approaches (Nahmias et al. 2013; Selmi et al. 2014, 2015; Hurtado and Javaloyes 2015) – including those based on optical pumping – fall under this category. While distinct, the output wavelength often has a stringent relationship with the input wavelength. For example, excitable micropillar lasers (Selmi et al. 2014; Barbay et al. 2011) are carefully designed to support one input mode with a node coincident with an antinode of the lasing mode. In cases where the input is also used as a pump (Shastri et al. 2016), the input wavelength must be shorter than that of the output in order to achieve carrier population inversion.

WDM networking introduces wavelength constraints that conflict with the ones inherent to optical injection. One approach for networking optically injected devices is to attempt to separate these wavelength constraints. In early work on neuromorphic photonics in fiber, this was accomplished with charge-carrier-mediated cross-gain modulation (XGM) in an SOA (Rosenbluth et al. 2009; Kravtsov et al. 2011; Fok et al. 2011).

O/E/O PNNs

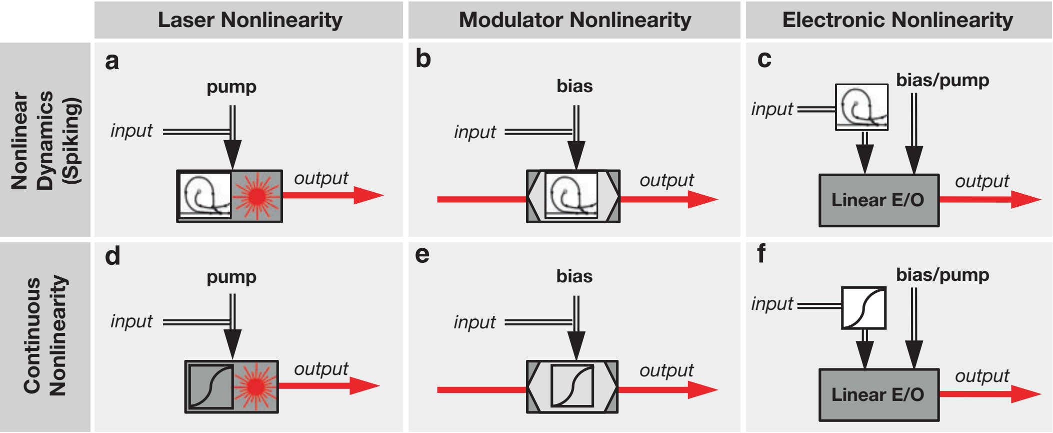
In this kind of PNN, the O/E subcircuit is responsible for the weighted addition functionality, whereas the E/O is responsible for the nonlinearity (Fig. 11c). Each subcircuit can therefore be analyzed independently. The analysis of an O/E WDM weighted addition circuit is deferred to a later section (photonic neural networks).

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 12 classifies six different ways 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 $\dot{x} = g\left(\vec{x}, u\right); \dot{y} = h\left(\vec{x}, y, u\right)$, 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, relaxation oscillations, etc. covered, for example, in Izhikivich (2007).

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 (Fig. 12). Figure 12a illustrates spiking lasers, which are detailed in the next section and offer perhaps the most promise in terms of garnering the full advantage of recent theoretical results on spike processing efficiency and expressiveness. Figure 12b is a spiking modulator. The work in Van Vaerenbergh et al. (2012) might be adapted to fit this classification; however, to the authors' knowledge, an ultrafast spiking modulator remains to be experimentally demonstrated. Figure 12c illustrates a purely electronic approach to nonlinear neural behavior. Linear E/O could be done by either a modulator or directly driven laser. This class could encompass interesting intersections with efficient analog electronic neurons in silicon (Indiveri et al. 2011; Pickett et al. 2013). 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 12d describes a laser with continuous nonlinearity, an instantiation of which was recently demonstrated in Nahmias et al. (2016). Figure 12e shows a modulator with continuous nonlinearity; the first demonstration of which in a PNN and recurrent network is presented in Tait et al. (2017). The pros and cons between the schemes in Fig. 12d, e are the same ones brought up by the on-chip versus off-chip light source debate, currently underway in the silicon photonic community. On-chip sources could provide advantageous energy scaling (Heck and Bowers 2014), but they require the introduction of exotic materials to the silicon photonic process to provide optical gain. Active research in this area has the goal of making source co-integration feasible (Liang and Bowers 2010a; Roelkens et al. 2010). An opposing school of thought argues that on-chip sources are still a nascent technology (Vlasov 2012). While fiber-to-chip coupling presents practical issues (Barwicz et al. 2016), discrete laser sources are cheap and well understood. Furthermore, on-chip lasers dissipate large amounts of power (Sysak et al. 2011), the full implications of which may complicate system design (Vlasov 2012). 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 give more attention to this strategy. Moreover, although the E/O part of the PNN can involve any kind of nonlinearity (Fig. 12), 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.



Neuromorphic Photonics, Principles of, Fig. 12 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 Nahmias et al. (2016). (**e**) Continuous modulator neuron, as demonstrated in Tait et al. (2017). (**f**) Continuous purely electronic nonlinearity with optical output (Reproduced from Ferreira de Lima et al. 2017). Licensed under Creative Commons Attribution-NonCommercial-NoDerivatives License. (CC BY-NC-ND)

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 Fig. 10. In the next section, we review recent developments of semiconductor excitable lasers that emulate spiking neural behavior, and, following that, we discuss a scalable WDM networking scheme.

Excitable/Spiking 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 of a *refractory period* during which it is temporarily less likely to emit another spike.

Analogy to Leaky Integrate-and-Fire Model

Excitable behavior can be realized near the threshold of a passively Q-switched two-section laser with saturable absorber (SA). Figure 13a, b shows an example of integrated design in a hybrid photonic 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 a saturable absorber region. The laser emits light along the waveguide structure into a passive silicon network. Figure 13c–e shows experimental data from a fiber ring laser prototype, demonstrating key properties of excitability.

In general, the dynamics of a two-section laser composed of a gain section and a saturable absorber (SA) can be described by the Yamada model (Eqs. 2, 3, and 4) (Yamada 1993). This 3D dynamical system, in its simplest form, can be described with the following undimensionalized equations (Nahmias et al. 2013; Barbay et al. 2011):

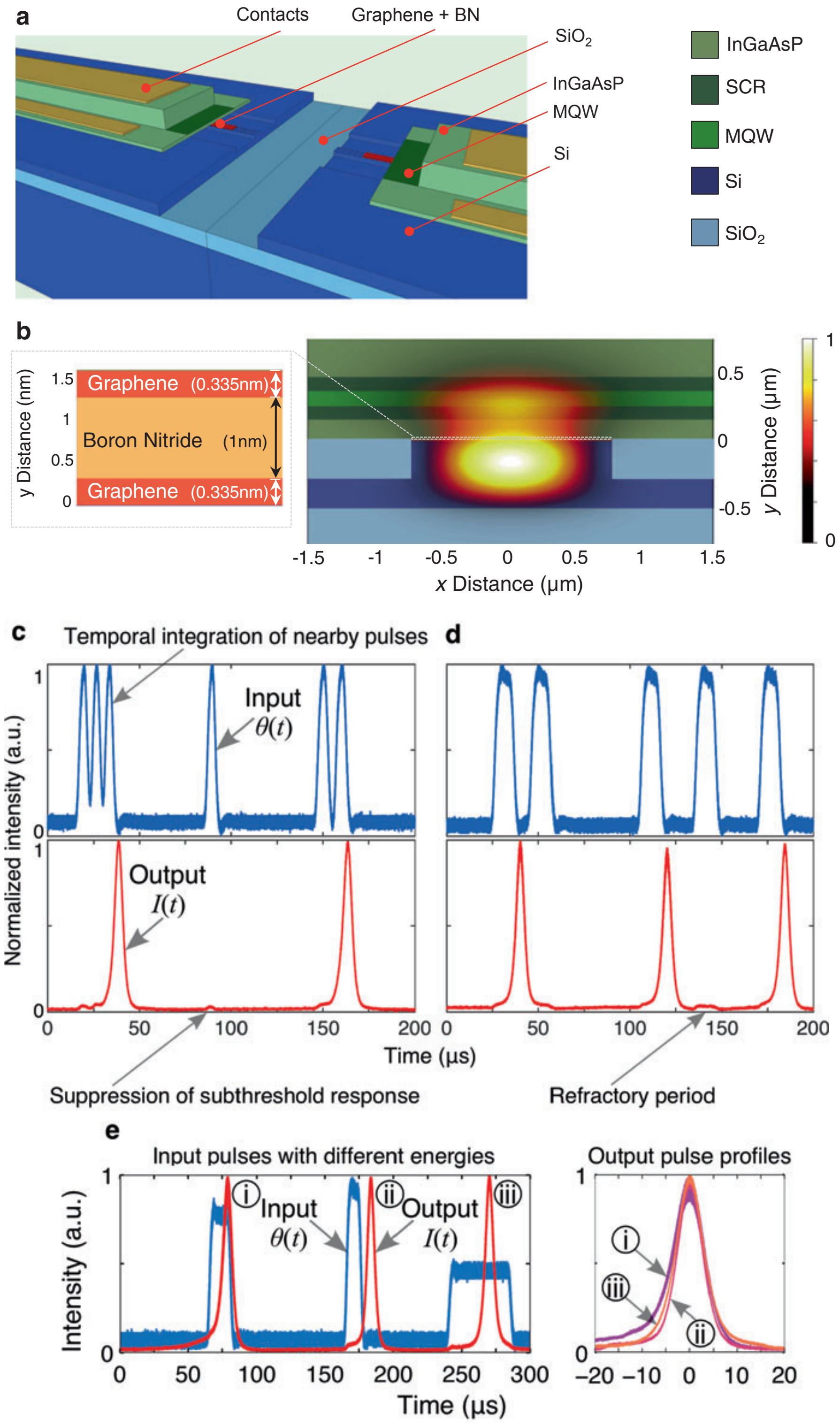


Fig. 13 (continued)

$$\frac{\mathrm{d}G(t)}{\mathrm{d}t} = \gamma_G[A - G(t) - G(t)I(t)] + \theta(t) \qquad (2)$$

$$\frac{\mathrm{d}Q(t)}{\mathrm{d}t} = \gamma_Q[B - Q(t) - aQ(t)I(t) \qquad (3)$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \gamma_I[G(t) - Q(t) - 1]I(t) + \varepsilon f(G), \qquad (4)$$

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 Nahmias et al. (2013) that the near-threshold dynamics of the laser described can be approximated to Eq. 5:

$$\begin{aligned}&\frac{\mathrm{d}G(t)}{\mathrm{d}t} = -\gamma_G(G(t) - A) + \theta(t);\\ &\text{if } G(t) > G_{\text{thresh}}\mathbb{Z}\\ &\text{release a pulse, and set } G(t) \to G_{\text{reset}},\end{aligned} \qquad (5)$$

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.

One can note the striking similarity to the LIF model in Eq. 1: setting the variables $\gamma_G = 1/R_m C_m$, $A = V_L$, $\theta(t) = I_{app}(t)/R_m C_m$, 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 $R_m C_m$ constant of the order of ms, the carrier dynamics in lasers are as fast as ns. Although this form of excitability was found in two-section lasers, other device morphologies have also shown excitable dynamics. The advantage 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.

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, summarized in Table 2. We group them under the terminology *excitable lasers* for convenience, but exceptions are described in the caption of the table.

Neuromorphic Photonics, Principles of, Fig. 13 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 $\left|\vec{E}\right|^2$ 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. (2016). Licensed under Creative Commons Attribution License (CC BY))

Neuromorphic Photonics, Principles of, Table 2 Characteristics of recent excitable laser devices. Note that this table does not have a one-to-one correspondence to Fig. 12, because some of them are not E/O devices. However, we observe that devices A, D, and F belong to the category Fig. 12a and device E resembles more closely to the category Fig. 12c

Device	Injection scheme	Pump	Excitable dynamics	References
A. Two-section gain and SA	Electrical	Electrical	Stimulated emission	(Nahmias et al. 2013, 2015; Selmi et al. 2014, 2015; Shastri et al. 2014, 2015, 2016; Barbay et al. 2011; Spühler et al. 1999; Dubbeldam and Krauskopf 1999; Dubbeldam et al. 1999; Larotonda et al. 2002; Elsass et al. 2010)
B. Semiconductor ring laser	Coherent optical	Electrical	Optical interference	(Coomans et al. 2010, 2011, 2013; Van Vaerenbergh et al. 2012; Gelens et al. 2010)
C. Microdisk laser	Coherent optical	Electrical	Optical interference	(Van Vaerenbergh et al. 2013; Alexander et al. 2013)
D. 2D photonic crystal nanocavity[a]	Electrical	Electrical	Thermal	(Brunstein et al. 2012; Yacomotti et al. 2006a, b)
E. Resonant tunneling diode photodetector and laser diode[b]	Electrical or incoherent optical	Electrical	Electrical tunneling	(Romeira et al. 2013; 2016)
F. Injection-locked semiconductor laser with delayed feedback	Electrical	Electrical	Optical interference	(Kelleher et al. 2010, 2011; Garbin et al. 2014, 2015; Goulding et al. 2007; Wieczorek et al. 1999, 2002, 2005; Barland et al. 2003; Marino and Balle 2005; Turconi et al. 2013)
G. Semiconductor lasers with optical feedback	Incoherent optical	Electrical	Stimulated emission	(Aragoneses et al. 2014; Sorrentino et al. 2015; Giudici et al. 1997; Yacomotti et al. 1999; Giacomelli et al. 2000; Heil et al. 2001; Wünsche et al. 2001)
H. Polarization switching VCSELs	Coherent optical	Optical	Optical interference	(Hurtado and Javaloyes 2015; Hurtado et al. 2010, 2012)

[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

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 Fig. 11) 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 or O/E/O PNNs. Some of them (B, C, H) operate free of electrical injection, meaning that bits of information remain elegantly encoded in optical carriers. However, as we have pointed out, 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.

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 than the others in which it does not require exchange of energy between charge carriers 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 Prucnal et al. (2016), but a quantitative comparison between performance metrics of lasers in Table 2 is still called for. Qualitatively, however, excitable lasers can simultaneously borrow key properties of electronic transistors, such as thresholding and cascadability.

In addition to individual laser excitability, there have been several demonstrations of simple processing circuits. Temporal pattern recognition (Shastri et al. 2016) and stable recurrent memory (Shastri et al. 2016; Romeira et al. 2016; Garbin et al. 2014) are essential toy circuits that demonstrate basic aspects of network compatibility.

Photonic Neural Network Architecture

Isomorphism to Biological Spiking Neuron

Neurons only have computational capabilities if they are in a network. Therefore, an excitable laser (or spiking laser) can only be viewed as a neuron candidate if it is contained in a PNN (Fig. 14). 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. Earlier, 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. (2014b) 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.

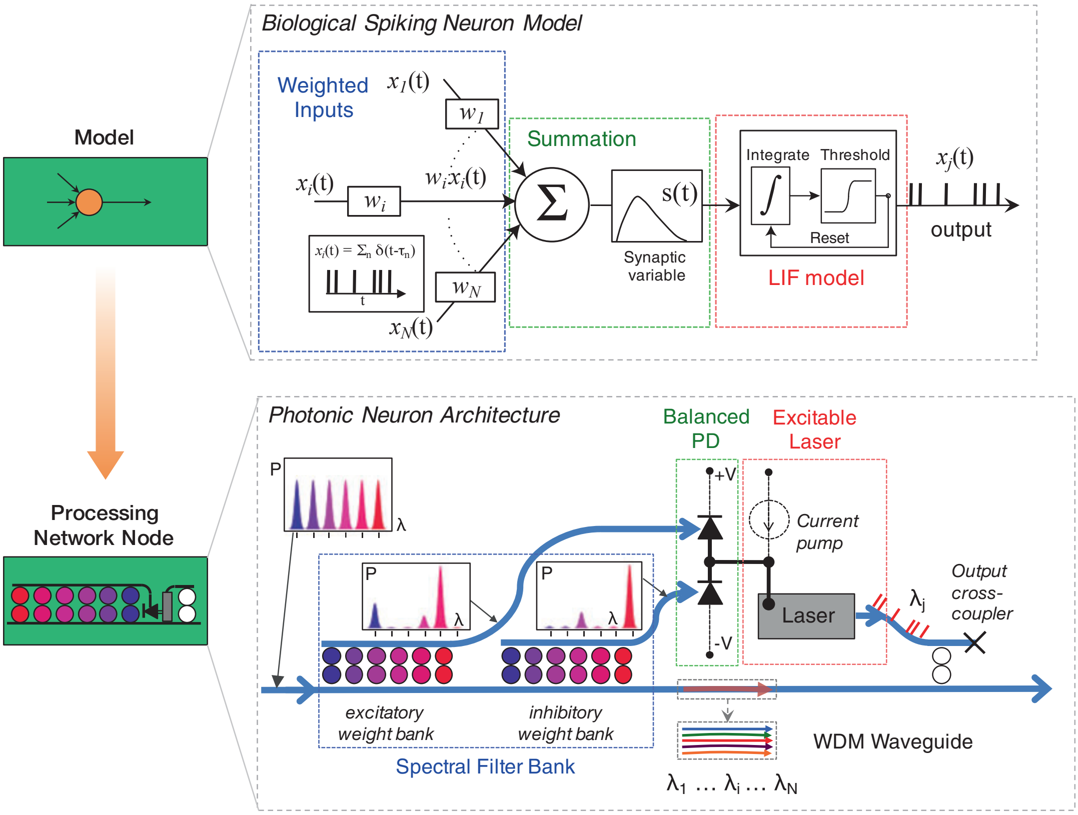
A spiking and/or analog photonic network consists of three aspects: a protocol, a node that abides by that protocol (the PNN), and a network medium that supports multiple connections between these nodes. This section will begin with broadcast-and-weight as a WDM protocol in which many signals can coexist 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 (Fig. 15). We will summarize experimental investigations of MRR weight banks.

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) (Preston et al. 2011). In fiber communication networks, a WDM protocol called broadcast-and-*select* has been used for decades to create many potential connections between communication nodes (Ramaswami 1993). 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 (Fig. 15b) 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 a MRR – in simple words, a waveguide bent

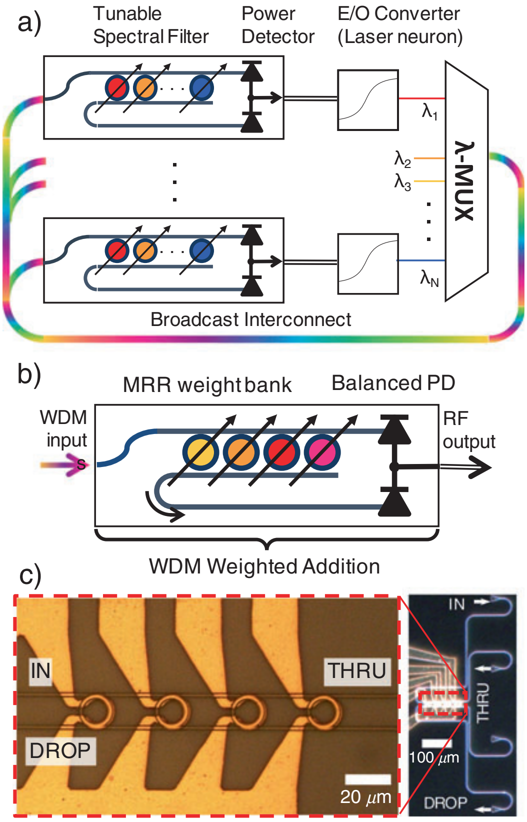


Neuromorphic Photonics, Principles of, Fig. 14 Isomorphism between photonic neuron module (i.e., a PNN) to a biological spiking neuron model. Top: depiction of a single unit neural network model. Inputs xi are weighted and summed. The result $\sum_i w_i x_i$ experiences a nonlinear function. Bottom: Possible physical implementation of a PNN. A WDM signal is incident on a bank of filters (excitatory and inhibitory) which are created using a series of microring filters that apply a series of weights. The resulting signal is incident on a balanced photodetector which applies a summation operation and drives an excitable laser with a current signal. The resulting laser outputs at a specified wavelength which is subsequently coupled back into the broadcast interconnect (see Fig. 15) (Adapted with permission from Tait et al. *J. Lightwave Technol.* **32**, 4029–4041 (2014) (Tait et al. 2014). Copyright 2014 Optical Society of America)

back on itself to create an interference condition. The MRR resonance wavelength can be tuned thermally (as in Fig. 15c) 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.

Controlling Photonic Weight Banks

Sensitivity to fabrication variations, thermal fluctuations, and thermal cross talk has made MRR control an important topic for WDM demultiplexers (Klein et al. 2005), high-order filters (Mak et al. 2015), modulators (Cox et al. 2014), and delay lines (Cardenas et al. 2010). 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 other hand, an MRR weight must be biased at arbitrary points in the filter roll-off region in order to multiply an optical signal by a continuous range of weight values. Feedback control approaches are well-suited to MRR demultiplexer and modulator control (DeRose et al. 2010; Jayatilleka et al. 2015), but these approaches rely on having a reference signal with consistent average power. In analog



Neuromorphic Photonics, Principles of, Fig. 15 (**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. (**b**) Tunable spectral filter constructed using microring resonator (MRR) weight bank. Tuning MRRs between on- and off-resonance switches a continuous fraction of optical power between drop and through ports. A balanced photodetector (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 (Reproduced with permission from Tait et al. (2016a). Copyright 2016 Optical Society of America)

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.

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 (Akopyan et al. 2015). In Tait et al. (2016b), 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 versus 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 Tait et al. (2016b) indicates a dynamic range of the weight controller of 9.2 dB, in other words, an equivalent digital resolution of 3.1 bits.

Multichannel Control Accuracy and Precision

Another crucial feature of an MRR weight bank is simultaneous control of all channels. When sources of cross talk 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 scales exponentially with the channel count, since the dimension of the range grows with channel count. Simultaneous control in the presence of cross talk therefore motivates model-based calibration approaches.

Model-based, as opposed to interpolation-based, calibration involves parameterized models for cross talk inducing effects. The predominant sources of cross talk are thermal leakage between nearby integrated heaters and, in a lab setup, 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 cross talk occurs when heat generated at a particular heater affects the temperature of neighboring devices (see Fig. 15c). 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. Tait et al. (2016a) describes 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-point 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. While optimal weight resolution is still a topic of discussion in the neuromorphic electronics community (Hasler and Marr 2013), several state-of-the-art architectures with dedicated weight hardware have settled on 4-bit resolution (Akopyan et al. 2015; Friedmann et al. 2013).

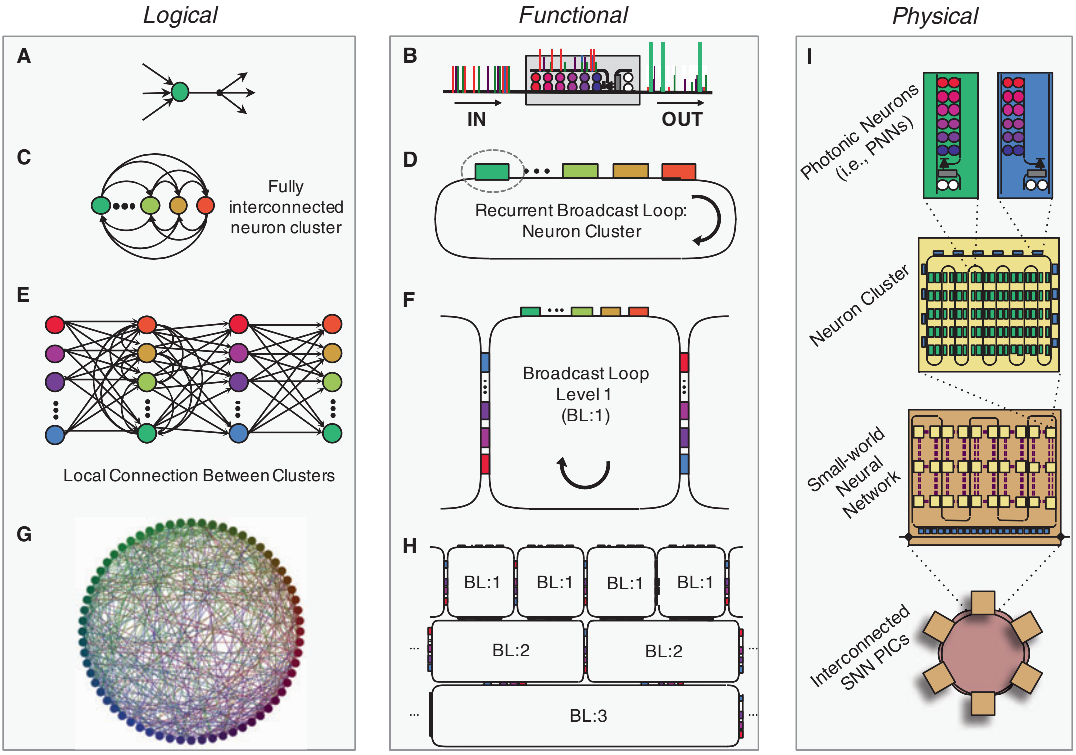
Scalability with Photonic Weight Banks

Engineering analysis and design relies 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 aspect 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 was 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 (Tait et al. 2016c).

In conventional analyses of MRR devices for multiplexing, demultiplexing, and modulating WDM signals, the trade-off that limits channel spacing is interchannel cross talk (Preston et al. 2011; Jayatilleka et al. 2016). However, unlike MRR demultiplexers where each channel is coupled to a distinct waveguide output (Klein et al. 2005), 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 cross talk between signals breaks down (Fig. 15b). 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 weight neighboring signals independently. As detailed in Tait et al. (2016c), Tait et al. (1) quantify this effect as a power penalty by including a notion of tuning *range* with the cross-weight penalty metric and (2) perform a channel density analysis by deriving the scalability of weight banks that use microresonators of a particular finesse.

In summary, WDM channel spacing, $\delta\omega$, can be used to determine the maximum channel count given a resonator finesse. While 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 linewidths depending on bus length. High finesse silicon MRRs, such as shown in Xu et al. (2008)


Logical
Functional
Physical
A
B
IN
OUT
C
Fully interconnected neuron cluster
D
Recurrent Broadcast Loop: Neuron Cluster
E
Local Connection Between Clusters
F
Broadcast Loop Level 1 (BL:1)
G
H
BL:1
BL:1
BL:1
BL:1
BL:2
BL:2
BL:3
I
Photonic Neurons (i.e., PNNs)
Neuron Cluster
Small-world Neural Network
Interconnected SNN PICs

Neuromorphic Photonics, Principles of, Fig. 16 Spectrum reuse strategy. Panels are organized into rows at different scales (core, cluster, chip, and multi-chip) and into columns at three different views (logical, functional, and physical). (**a**, **b**) Depicts the model and photonic implementation of a neuron as a processing network node (PNN) as detailed in Fig. 14. (**c**, **d**) Fully interconnected network by attaching PNNs to a broadcast loop (BL) waveguide. (**e**, **f**) A slightly modified PNN can transfer information from one BL to another. (**g**, **h**) Hierarchical organization of the waveguide broadcast architecture showing a scalable modular structure. (i) Using this scheme, neuron count in one chip is only limited by footprint, but photonic integrated circuits (PICs) can be further interconnected in an optical fiber network

(finesse = 368) and (Biberman et al. 2012) (finesse = 540), could support 108 and 148 channels, respectively. Other types of resonators in silicon, such as elliptical microdisks (Xiong et al. 2011) (finesse = 440) and traveling-wave microresonators (Soltani et al. 2010) (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 Tait et al. (2016a), 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. Tait et al. (2016c) arrived at a scaling limit of 148 channels for a 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 (Fig. 16) proposed in Tait et al. (2014b), by tailoring interference within MRR weight banks as discussed in Tait et al. (2016c), 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 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.

Neuromorphic Platform Comparison

As stated earlier, the neuromorphic computing community has been making vigorous efforts toward large-scale spiking neuromorphic hardware, e.g., Heidelberg HICANN chip via the FACETS/BrainScaleS projects (Schemmel et al. 2010), IBM TrueNorth via the DARPA SyNAPSE program (Merolla et al. 2014), Stanford University's Neurogrid (Benjamin et al. 2014), and SpiNNaker (Furber et al. 2014) (Fig. 4). Many researchers are concentrating their efforts toward the long-term technical potential and functional capability of the hardware compared to standard digital computers. One of the main drivers for the community is computational power efficiency (Hasler and Marr 2013): digital CPUs are reaching a power efficiency wall with the current von Neumann architecture, but hardware neural networks, in which memory and instructions are simplified and colocated, offer to overcome this barrier.

These projects also aimed at simulating large-scale spiking neural networks, with the goal of simulating subcircuits of the human cortex, at a biological timescale (<1 kHz). The HICANN chip, exceptionally, is designed to be accelerated at about 10,000 times respective to biological timescales and features analog synapses and realistic neural spiking behaviors (Schemmel et al. 2010). It pays the price of huge power consumption: 800 W for a wafer-scale system containing 180,000 neurons (Benjamin et al. 2014). In contrast, TrueNorth aimed for large-scale, efficient networks optimized for biologically plausible tasks, such as machine vision, but with a simplified neural model (Merolla et al. 2014). Indeed, it contained a total of 1 M neurons and 256 M synapses per chip, consuming only 63 mW of power (Merolla et al. 2014). The Neurogrid board also aimed for scalability and efficiency, consuming 5 W.

also with 1 M neurons and about four billion synapses, but it kept greater biological fidelity to the mammalian cortex. The SpiNNaker computer is designed to simulate scalably large and versatile networks using arrays of chips with 18 ARM968 processing cores each: unlike the other three technologies, the number of synapses per neuron, or even the number of neurons, is not fixed and can be dynamically reprogrammed (Furber et al. 2014). The demonstrated system has 48 chips interconnected in a PCB, but it can be scaled up to 1200 interconnected PCBs, totalling a 72 kW peak power consumption.

We have recently produced a quantitative comparison between the aforementioned electronic and photonic neuromorphic hardware architectures (Prucnal and Shastri 2017). In order to compare these processors with one another, we reintroduce the multiply-accumulate (MAC) operation that typically bottlenecks complex computations.

The MAC operation takes the following form: $a \leftarrow a + (w \times x)$. It includes both a product (i.e., x is multiplied by the "weight" w) and an addition (the result is accumulated to variable a). In neural network models, inputs are combined via a weighted sum of the form $\Sigma\ w_i x_i$. The result is then input into some nonlinear scalar function $f\{x\}$ which can range from a simple sigmoid function to a complex nonlinear dynamical system with hysteresis, depending on the complexity of the neuron being modeled. The weighted sum can be broken down into a series of MAC operations of the form $a_i = a_{i-1} + w_i x_i$ for $i = 1 \ldots M$. Each neuron requires M parallel MAC operations per time step Δt (in a given bit period τ, determined by the signal bandwidth capacity) or one operation per synapse, where M refers to the number of inputs for a given node. Thus, a hardware neural network can be characterized with $M \times N$ MAC operations per time step Δt (i.e., quadratic scaling of MAC operations), where N is the number of neurons in the network.

The nonlinear function $f\ \{x\}$ also takes up computational resources, but since this operation scales with N rather than $M \times N$, it does not represent the most costly operation. Therefore, as the size of the network N grows large, MACs – i.e., "synaptic computations" – become the most burdensome hardware bottlenecks in neural networks (Hasler and Marr 2013).

For consistency, we compare architectures that have similar functionality: we limit ourselves to fully reconfigurable systems of spiking neural networks. For the photonically enhanced system, we studied an optoelectronic neural network with PNNs instantiated within the hybrid silicon/III–V platform (Nahmias et al. 2015; Ferreira de Lima et al. 2016). 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. We also refer the reader to a more detailed discussion of spiking electronic neuromorphic hardware in Liu et al. (2015) and an overview of current spiking and nonspiking hardware in Nawrocki et al. (2016).

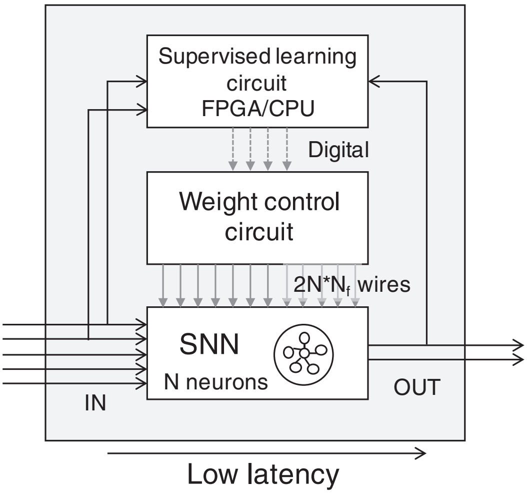
Results are summarized in Table 1. The most striking figure is the number of operations per second, which exceeds electronic platforms by three orders of magnitude, compared to the analog/digital accelerated HICANN and six orders of magnitude compared to the others which are purely digital implementations. This stems from both the high bandwidths and low latencies possible with photonic signals. The optoelectronic approach is able to achieve such energy efficiency at high speeds because power is mainly consumed statically by the lasers, while the passive filters have low leakage current. This contrasts with 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, since 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, even though 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 electronics systems presented here.

Future Directions

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 elements of a reconfigurable photonic hardware that can emulate spiking neural networks 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 say (Chrostowski and Hochberg 2015) 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. It is reasonable to expect that neuromorphic photonic coprocessors (Fig. 17) can be fabricated and packaged using fabless services for near-term research and long-term volume production.

Applications for neuromorphic photonic processors can be grouped into two categories: (1) a front-end stage for RF systems and data centers and (2) ultrafast processing for specialized fast applications (Prucnal and Shastri 2017). The first category utilizes the low-latency, parallelism, and energy efficiency properties of photonics to alleviate the throughput of RF systems, e.g., 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.

Neuromorphic photonic processors join a class of photonic hardware accelerators designed to



Neuromorphic Photonics, Principles of, Fig. 17 Diagram description of a fully packaged neuromorphic processor. While two layers of electronics provide reconfigurability, the photonic spiking neural network permits low-latency functionality. N_f: fan-in of each neuron (Reproduced from Ferreira de Lima et al. 2017). Licensed under Creative Commons AttributionNonCommercial-NoDerivatives License. (CC BY-NC-ND))

assist in acquisition, feature extraction, and storage of wideband waveforms (Jalali and Mahjoubfar 2015). These accelerators manipulate the spectrotemporal of a wideband signal, a task difficult to accomplish in analog electronics over broad bandwidth and with low loss.

Real-Time Radio-Frequency Processing

High-volume data applications, including streaming video and cloud services, will continue to push the telecommunications industry to build better, high-bandwidth systems. Data traffic on some mobile networks alone has increased by over 6000% (Index 2015). This has motivated the exploration of more efficient usage of spectral resources (Akyildiz et al. 2006). Although RF integrated circuits (RFICs) have been researched for applications such as duplex processing (Lee 2003; Razavi 2000) or control of beamforming antennas (Yu et al. 2011; Razavi 2009), the requirement for impedance-matched transmission lines greatly increases device and interconnect footprint, limiting the overall complexity of each chip. Photonics provides a solution to these fundamental limitations: optical waveguides can support large bandwidths (~100 s of THz) with high information density and low cross talk between multiplexed channels. By using techniques such as WDM, a large number of multi-GHz channels can exist within the same optical waveguide. As a result, the number of virtual channels can greatly exceed the number of physical waveguides, allowing for the formation of complex processing circuits without a significant hardware overhead.

After some initial front-end processing (i.e., heterodyning and amplification), most radio transceiver systems are processed by either digital signal processors (DSP) 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 the ~GHz. 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 re-configurability. 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 (Larsson et al. 2014; Gesbert et al. 2003) – are especially susceptible to this bottleneck and may require a radically new solution.

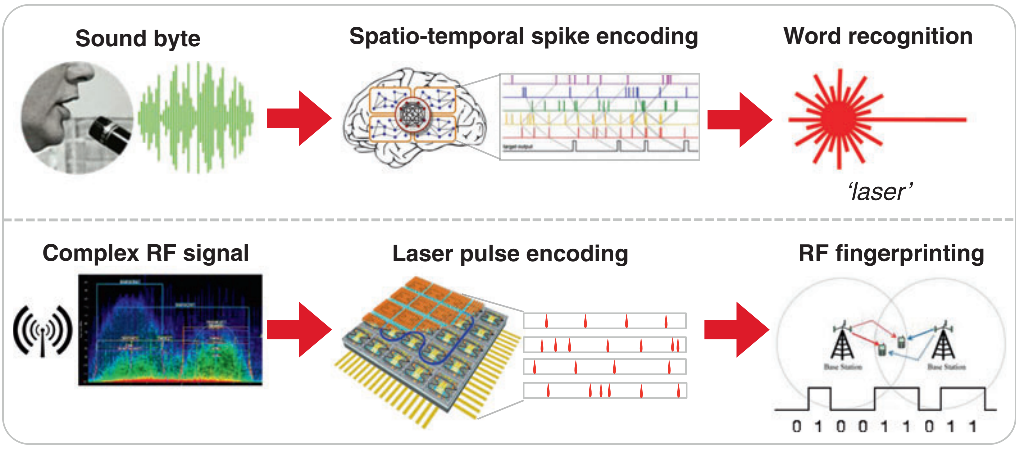
Adding a photonic processing chip to the front of a radio transceiver would allow very complex operations to be performed in real time, which can significantly offload electronic post-processing and provide a technology to make faster, more relevant RF decisions on the fly. Massive MIMO systems based 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 (Larsson et al. 2014; Hansen 2009). 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.

As spread spectrum, adaptive RF transceivers become more widespread in future telecommunication systems, the scalability of the photonic approach could provide significant processing advantages. Its high bandwidth, low latency, and high throughput would be especially useful in an ultra-wideband (UWB) radio system, in which it could sample from many frequencies and directions simultaneously to scan for spectral opportunities and make a decision quickly and efficiently. Pairing this technology with an FPGA or electronic ASIC controller would enable the implementation of adaptive optimization and learning algorithms in real time for ultrafast cognitive radio applications.

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. Since one of the most salient advantages of a photonic approach is its low time of flight (~ps) between communicating processors, the convergence rates can be significantly improved by implementing them on a photonic platform. A large class of problems that can be solved iteratively includes *linear and nonlinear programming problems.* These methods seek to minimize some objective function $E\left(\vec{x}\right)$ of real variables in $\vec{x}$ subject to a series of constraints represented by equalities or inequalities, i.e., $g\left(\vec{x}\right) \leq 0$ and $h\left(\vec{x}\right) = 0$. Applications in telecommunications, aerospace, and financial industries can be described in this basic framework, including optimal portfolio trading strategies, control of machinery/actuators, and allocation of resources and jobs in online servers. Using a photonic approach, 100-variable problems could converge in less than $\sim$100 ns, which could be useful in the control of very fast dynamical systems (i.e., actuators) or in the creation of low-latency optimization routines in data-intensive environments.

Mathematical optimization problems can be grouped into *linear* and *nonlinear* optimization problems. Nonlinear optimization problems are often difficult to solve and sometimes involve exotic techniques such as genetic algorithms or particle swarm optimization. Nonlinear optimization problems, however, are nonetheless *quadratic* to second order around the local vicinity of the optimum. Therefore, quadratic programming (QP) – which finds the minima/maxima quadratic functions of variables subject to linear constraints (Lendaris et al. 1999) – becomes an effective first pass at such problems and can be applied to a wide array of applications. For example, many machine learning problems, such as support vector machine (SVM) training and least-squares regression, can be reformulated in terms of a QP problem. In addition, computational problems such as model predictive control (MPC), an optimal nonlinear control algorithm, or compressive sampling, a method for sampling at rates below the Nyquist without loss of information via the characterization of sparsity in incoming data, are examples of QP problems. Together, these applications represent some of the most effective yet generalized tools for acquiring and processing information and using the results to control systems. QP is an NP hard problem in the number of variables, which means that conventional digital computers must either be limited to solving quadratic programs of very few variables or to applications where computation time is noncritical. This is reflected in industrial applications of QP solvers. MPC is used in the chemical industry to control chemical processing plants, where reaction timescales can be made very long, and in the finance industry to control long-term portfolio optimization. The application of MPC to faster systems, therefore, relies on new ways of finding faster solutions to QP problems (Jerez et al. 2011). In machine learning, many algorithms (such as SVM) require offline training because of the computational complexity of QP but would be much more effective if they could be trained online (Fig. 18).

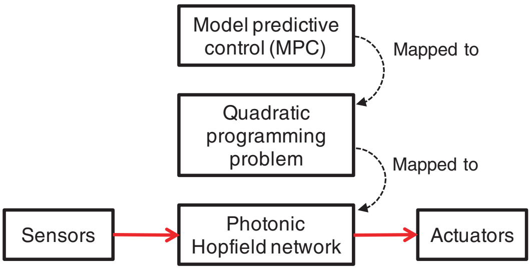


Neuromorphic Photonics, Principles of, Fig. 18 A vision for a brain-inspired laser neural network for enhanced RF communication. Top: spoken words are transformed into spatiotemporal "spike" (event) patterns by biological neurons. A pattern recognition neuron is sensitive to a specific spike fingerprint and releases its own spike only if it occurs, shown in "target output." Bottom: in analogy with audio waveforms, a much faster system (~GHz) based on excitable lasers could operate directly on RF waveforms. Applying operations at the front-end of RF transceivers could offload complex signal processing operations to a photonic chip and address bandwidth and latency limitations of current FPGA and DSP solutions (The spike-coded pattern is reproduced from Tapson et al. 2013). Licensed under Creative Commons Attribution License (CC BY))

While Hopfield showed that Hopfield networks are able to solve quadratic optimization problems quickly over 25 years ago (Tank and Hopfield 1986), Hopfield quadratic optimizers are uncommon today. This is largely due to the high connectivity between neurons that neural networks require (n^2 connections for n neurons). In an electronic circuit, as the number of connections increases, the bandwidth at which the system can operate without being subject to cross talk between connections and other issues decreases (Tait et al. 2014b). This creates an undesirable trade-off between neuron speed and neural network size. Photonic neural networks have several advantages over their electrical counterparts. Most importantly, the connectivity concerns prevalent in electronic neurons are significantly ameliorated by using light as a communication medium (Tait et al. 2014b). WDM allows for hundreds of high-bandwidth signals to flow through a single optical waveguide. Moreover, the analog computational bandwidth of a photonic neuron (as designed in Tait et al. (2016b)) lies in the picosecond to femtosecond timescale (Nahmias et al. 2015). For a Hopfield quadratic optimizer, this means that a photonic implementation (Fig. 19) can simultaneously have large dimensionality and a fast convergence time to the minimum. 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 (Keviczky and Balas 2006).

Summary and Conclusion

Photonics has revolutionized information communication, while electronics, in parallel, has dominated information processing. Recently, there has been a determined exploration of the unifying boundaries between photonics and electronics on the same substrate, driven in part by Moore's law approaching its long-anticipated end. For example, the computational efficiency for digital processing has leveled off around 100 pJ per MAC. As a result, there has been a widening gap between today's computational efficiency and the next-generation needs, such as big data applications which require advanced pattern matching and real-time analysis. This, in turn, has led to expeditious advances in (1) emerging devices that


Model predictive control (MPC)
Mapped to
Quadratic programming problem
Mapped to
Sensors
Photonic Hopfield network
Actuators

Neuromorphic Photonics, Principles of, Fig. 19 Employing a photonic neural network for a model predictive control problem by solving a quadratic programming problem with a Hopfield network

are called “beyond CMOS” or “More-than-Moore”; (2) novel processing or unconventional computing architectures called “beyond von Neumann” that are brain-inspired, i.e., neuromorphic; and (3) CMOS-compatible photonic interconnect technologies. Collectively, these research endeavors have given rise to the field of neuromorphic photonics (Fig. 1). Emerging photonic hardware platforms have the potential to vastly exceed the capabilities of electronics by combining ultrafast operation, moderate complexity, and full programmability, extending the bounds of computing for applications such as navigation control on hypersonic aircrafts and real-time analysis of the RF spectrum.

In this entry, we discussed the current progress and requirements of such a platform. In a photonic spike processor, information is encoded as events in the temporal and spatial domains of spikes (or optical pulses). This hybrid coding scheme is digital in amplitude but analog in time and benefits from the bandwidth efficiency of analog processing and the robustness to noise of digital communication. Optical pulses are received, processed, and generated by certain class of semiconductor devices that exhibit excitability – a nonlinear dynamical mechanism underlying all-or-none responses to small perturbations. Optoelectronic devices operating in the excitable regime are dynamically analogous to a biophysical neuron, but roughly eight orders of magnitude faster. We dubbed these devices as “photonic neurons” or “laser neurons.” The field is now reaching a critical juncture where there is a shift from studying single photonic neurons to studying interconnected networks of such devices. A recently proposed on-chip networking architecture called broadcast-and-weight could support massively parallel (*all-to-all*) interconnection between excitable devices using wavelength division multiplexing.

A hybrid III–V/Si photonic platform is a candidate for an integrated hardware platform. III–V compound semiconductor technology, such as indium phosphide (InP) and gallium arsenide (GaAs), is at the forefront of providing *active* elements like lasers, amplifiers, and detectors. Silicon, in parallel, brings compatibility with CMOS fabrication processes and low-loss *passive* components like waveguides and resonators. Scalable and fully reconfigurable networks of excitable lasers can be implemented in the silicon photonic layer of modern hybrid integration platforms, in which spiking lasers in a bonded InP layer are densely interconnected through a silicon layer. Such a photonic spike processor will potentially be able to support several thousand interconnected devices. It is predicted that such a chip would have a computational efficiency of 260 fJ per MAC, which surpasses the energy efficiency wall by two orders of magnitude while operating at high speeds (i.e., signal bandwidths 10 GHz). The emerging field of photonic spike processors has received tremendous interest and continues to develop as photonic integrated circuits increase in performance and scale. As novel applications requiring real-time, ultrafast processing – such as the exploitation of the RF spectrum – become more demanding, we expect that these systems will find use in a variety of high-performance, time-critical environments.

Moving forward, we envision a tremendous interest in designing, building, and understanding photonic networks of excitable elements for ultrafast information processing, guided by the latest computational models of the brain. Successful implementation of a small-scale photonic spike processor could, in principle, provide the fundamental technology to build and study larger-scale brain-inspired networks based on laser excitability. Neuromorphic photonics is poised to usher in exciting new fields of inquiry and impactful enterprises of application.

Bibliography

Agrawal GP (2002) Fiber-optic communication systems. Wiley series in microwave and optical engineering (Wiley-interscience). Wiley, New York

Akopyan F, Sawada J, Cassidy A, Alvarez-Icaza R, Arthur J, Merolla P, Imam N, Nakamura Y, Datta P, Nam GJ, Taba B, Beakes M, Brezzo B, Kuang J, Manohar R, Risk W, Jackson B, Modha D (2015) IEEE Trans Comput Aided Des Integr Circuits Syst. TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip. 34:1537

Akyildiz IF, Lee WY, Vuran MC, Mohanty S (2006) Comput Netw 50:2127

Alexander K, Van Vaerenbergh T, Fiers M, Mechet P, Dambre J, Bienstman P (2013) Opt Express 21:26182

Aragoneses A, Perrone S, Sorrentino T, Torrent MC, Masoller C (2014) Sci Rep 4:4696 EP

Backus J (1978) Commun ACM 21:613

Barbay S, Kuszelewicz R, Yacomotti AM (2011) Opt Lett 36:4476

Barland S, Piro O, Giudici M, Tredicce JR, Balle S (2003) Phys Rev E 68:036209

Barwicz T, Taira Y, Lichoulas TW, Boyer N, Martin Y, Numata H, Nah JW, Takenobu S, Janta-Polczynski A, Kimbrell EL, Leidy R, Khater MH, Kamlapurkar S, Engelmann S, Vlasov YA, Fortier P (2016) IEEE J Sel Top Quantum Electron 22:455

Bengio Y, Courville A, Vincent P (2013) IEEE Trans Pattern Anal Mach Intell 35:1798

Benjamin B, Gao P, McQuinn E, Choudhary S, Chandrasekaran A, Bussat JM, Alvarez-Icaza R, Arthur J, Merolla P, Boahen K (2014) Proc IEEE 102:699

Bhalla US, Iyengar R (1999) Science 283:381

Biberman A, Shaw MJ, Timurdogan E, Wright JB, Watts MR (2012) Opt Lett 37:4236

Boahen K (2000) IEEE Trans Circuits Syst II, Analog Digit Signal Process 47:416

Borst A, Theunissen FE (1999) Nat Neurosci 2:947

Brunner D, Soriano MC, Mirasso CR, Fischer I (2013a) Nat Commun 4:1364

Brunner D, Soriano MC, Fischer I (2013b) IEEE Photon Technol Lett 25:1680

Brunstein M, Yacomotti AM, Sagnes I, Raineri F, Bigot L, Levenson A (2012) Phys Rev A 85:031803

Burgsteiner H (2005) On learning with recurrent spiking neural networks and their applications to robot control with real-world devices. PhD thesis, Graz University of Technology

Capmany J, Ortega B, Pastor D (2006) J Lightwave Technol 24:201

Cardenas J, Foster MA, Sherwood-Droz N, Poitras CB, Lira HLR, Zhang B, Gaeta AL, Khurgin JB, Morton P, Lipson M (2010) Opt Express 18:26525

Caulfield HJ, Dolev S (2010) Nat Photonics 4:261

Chrostowski L, Hochberg M (2015) Silicon photonics design: from devices to systems. Cambridge University Press, Cambridge

Coomans W (2012) Nonlinear dynamics in semiconductor ring lasers towards an integrated optical neuron. PhD thesis, Vrije Universiteit Brussel

Coomans W, Beri S, Sande GVD, Gelens L, Danckaert J (2010) Phys Rev A 81:033802

Coomans W, Gelens L, Beri S, Danckaert J, Van Der Sande G (2011) Phys Rev E 84:1

Coomans W, Van der Sande G, Gelens L (2013) Phys Rev A 88:033813

Cox JA, Lentine AL, Trotter DC, Starbuck AL (2014) Opt Express 22:11279

Crescenzi P, Goldman D, Papadimitriou C, Piccolboni A, Yannakakis M (1998) J Comput Biol 5:597

Dennard R, Rideout V, Bassous E, LeBlanc A (1974) IEEE J Solid State Circuits 9:256

DeRose CT, Watts MR, Trotter DC, Luck DL, Nielson GN, Young RW (2010) In: Conference on lasers and electro-optics 2010 (Optical Society of America), p CThJ3

Diesmann M, Gewaltig MO, Aertsen A (1999) Nature 402:529

Donges JF, Zou Y, Marwan N, Kurths J (2009) Eur Phys J Special Topics 174:157

Dubbeldam JLA, Krauskopf B (1999) Opt Commun 159:325

Dubbeldam JLA, Krauskopf B, Lenstra D (1999) Phys Rev E 60:6580

Duport F, Schneider B, Smerieri A, Haelterman M, Massar S (2012) Opt Express 20:22783

Duport F, Smerieri A, Akrout A, Haelterman M, Massar S (2016) Sci Rep 6:22381 EP

Eliasmith C (2005) Neural Comput 17:1276

Eliasmith C, Stewart TC, Choo X, Bekolay T, DeWolf T, Tang Y, Rasmussen D (2012) Science 338:1202

Elsass T, Gauthron K, Beaudoin G, Sagnes I, Kuszelewicz R, Barbay S (2010) Eur Phys J D 59:91

Esmaeilzadeh H, Blem E, St. Amant R, Sankaralingam K, Burger D (2012) IEEE Micro 32:122

Fang AW, Park H, Hao Kuo Y, Jones R, Cohen O, Liang D, Raday O, Sysak MN, Paniccia MJ, Bowers JE (2007) Mater Today 10:28

Ferreira de Lima T, Shastri BJ, Nahmias MA, Tait AN, Prucnal PR (2016) In: Summer topicals meeting series (SUM), 2016 (IEEE, 2016)

Ferreira de Lima T, Shastri BJ, Tait AN, Nahmias MA, Prucnal PR (2017) Nanophotonics 6:577

Fok MP, Deming H, Nahmias M, Rafidi N, Rosenbluth D, Tait A, Tian Y, Prucnal PR (2011) Opt Lett 36:19

Friedmann S, Frémaux N, Schemmel J, Gerstner W, Meier K (2013) Front Neurosci 7:160

Furber S, Galluppi F, Temple S, Plana L (2014) Proc IEEE 102:652
Garbin B, Goulding D, Hegarty SP, Huyet G, Kelleher B, Barland S (2014) Opt Lett 39:1254
Garbin B, Javaloyes J, Tissoni G, Barland S (2015) Nat Commun 6:5915 EP, 1409.6350
Gelens L, Mashal L, Beri S, Coomans W, Van der Sande G, Danckaert J, Verschaffelt G (2010) Phys Rev A 82:063841
Gesbert D, Shafi M, Shan Shiu D, Smith PJ, Naguib A (2003) IEEE J Sel Areas Commun 21:281
Giacomelli G, Giudici M, Balle S, Tredicce JR (2000) Phys Rev Lett 84:3298
Giudici M, Green C, Giacomelli G, Nespolo U, Tredicce JR (1997) Phys Rev E 55:6414
Goulding D, Hegarty SP, Rasskazov O, Melnik S, Hartnett M, Greene G, McInerney JG, Rachinskii D, Huyet G (2007) Phys Rev Lett 98:153903
Hansen RC (2009) Phased array antennas. Wiley, Hoboken, NJ
Hasler J, Marr B (2013) Front Neurosci 7:118
Heck M, Bowers J (2014) IEEE J Sel Top Quantum Electron 20:332
Heck M, Bauters J, Davenport M, Doylend J, Jain S, Kurczveil G, Srinivasan S, Tang Y, Bowers J (2013) IEEE J Sel Top Quantum Electron 19:6100117
Heil T, Fischer I, Elsäßer W, Gavrielides A (2001) Phys Rev Lett 87:243901
Hicke K, Escalona-Morán MA, Brunner D, Soriano MC, Fischer I, Mirasso CR (2013) IEEE J Sel Top Quantum Electron 19:1501610
Hidalgo CA, Klinger B, Barabási AL, Hausmann R (2007) Science 317:482
Hodgkin AL, Huxley AF (1952) J Physiol 117:500
Hopfield JJ (1982) Proc Natl Acad Sci 79:2554
Hurtado A, Javaloyes J (2015) Appl Phys Lett 107:241103
Hurtado A, Henning ID, Adams MJ (2010) Opt Express 18:25170
Hurtado A, Schires K, Henning ID, Adams MJ (2012) Appl Phys Lett 100:103703
Index CVN (2015) White Paper, February
Indiveri G, Linares-Barranco B, Hamilton T, van Schaik A, Etienne-Cummings R, Delbruck T, Liu SC, Dudek P, Häfliger P, Renaud S, Schemmel J, Cauwenberghs G, Arthur J, Hynna K, Folowosele F, SAÏGHI S, Serrano-Gotarredona T, Wijekoon J, Wang Y, Boahen K (2011) Front Neurosci 5:73
Izhikevich E (2003) IEEE Trans Neural Netw 14:1569
Izhikevich EM (2004) IEEE Trans Neural Netw 15:1063
Izhikivich EM (2007) Dynamical systems in neuroscience: the geometry of excitability and bursting. MIT Press, Cambridge
Jaeger H, Haas H (2004) Science 304:78
Jalali B, Fathpour S (2006) J Lightwave Technol 24:4600
Jalali B, Mahjoubfar A (2015) Proc IEEE 103:1071
Jayatilleka H, Murray K, Ángel Guillén-Torres M, Caverley M, Hu R, Jaeger NAF, Chrostowski L, Shekhar S (2015) Opt Express 23:25084
Jayatilleka H, Murray K, Caverley M, Jaeger N, Chrostowski L, Shekhar S (2016) J Lightwave Technol 34:2886
Jerez JL, Constantinides GA, Kerrigan EC (2011) ACM/SIGDA international symposium on field programmable gate arrays FPGA, Monterey, CA, p 209
Kelleher B, Bonatto C, Skoda P, Hegarty SP, Huyet G (2010) Phys Rev E 81:036204
Kelleher B, Bonatto C, Huyet G, Hegarty SP (2011) Phys Rev E 83:026207
Keviczky T, Balas GJ (2006) Control Eng Pract 14:1023
Keyes RW (1985) Opt Acta Int J Optics 32:525
Klein E, Geuzebroek D, Kelderman H, Sengo G, Baker N, Driessen A (2005) IEEE Photon Technol Lett 17:2358
Koomey J, Berard S, Sanchez M, Wong H (2011) IEEE Ann Hist Comput 33:46
Krauskopf B, Schneider K, Sieber J, Wieczorek S, Wolfrum M (2003) Opt Commun 215:367
Kravtsov K, Fok MP, Rosenbluth D, Prucnal PR (2011) Opt Express 19:2133
Kumar A, Rotter S, Aertsen A (2010) Nat Rev Neurosci 11:615
Larger L, Soriano MC, Brunner D, Appeltant L, Gutierrez JM, Pesquera L, Mirasso CR, Fischer I (2012) Opt Express 20:3241
Larotonda MA, Hnilo A, Mendez JM, Yacomotti AM (2002) Phys Rev A 65:033812
Larsson E, Edfors O, Tufvesson F, Marzetta T (2014) IEEE Commun Mag 52:186
Lee TH (2003) The design of CMOS radio-frequency integrated circuits, 2nd edn. Cambridge University Press. New York, NY
Lee WC, Hu C (2001) IEEE Trans Electron Devices 48:1366
Lendaris GG, Mathia K, Saeks R (1999) IEEE Trans Syst Man Cybern B (Cybern) 29:114
Liang D, Bowers JE (2010a) Nat Photonics 4:511
Liang D, Bowers JE (2010b) Nat Photonics 4:511
Liang D, Roelkens G, Baets R, Bowers JE (2010) Materials 3:1782
Liu SC, Delbruck T, Indiveri G, Whatley A, Douglas R (2015) Event-based neuromorphic systems. Wiley, Chichester
Maass W (1997) Neural Netw 10:1659
Maass W, Natschläger T, Markram H (2002) Neural Comput 14:2531
Mak J, Sacher W, Xue T, Mikkelsen J, Yong Z, Poon J (2015) IEEE J Quantum Electron 51:1
Marino F, Balle S (2005) Phys Rev Lett 94:094101
Markram H, Meier K, Lippert T, Grillner S, Frackowiak R, Dehaene S, Knoll A, Sompolinsky H, Verstreken K, DeFelipe J, Grant S, Changeux JP, Saria A (2011) Procedia Comput Sci 7:39
Marpaung D, Roeloffzen C, Heideman R, Leinse A, Sales S, Capmany J (2013) Laser Photonics Rev 7:506
Marr B, Degnan B, Hasler P, Anderson D (2013) IEEE Trans Very Large Scale Integr (VLSI) Syst 21:147

Martinenghi R, Rybalko S, Jacquot M, Chembo YK, Larger L (2012) Phys Rev Lett 108:244101

Mathur N (2002) Nature 419:573

Merkle RC (1989) Foresight Update 6

Merolla PA, Arthur JV, Alvarez-Icaza R, Cassidy AS, Sawada J, Akopyan F, Jackson BL, Imam N, Guo C, Nakamura Y, Brezzo B, Vo I, Esser SK, Appuswamy R, Taba B, Amir A, Flickner MD, Risk WP, Manohar R, Modha DS (2014) Science 345:668

Mesaritakis C, Papataxiarhis V, Syvridis D (2013) J Opt Soc Am B 30:3048

Miller DAB (2000) Proc IEEE 88:728

Modha DS, Ananthanarayanan R, Esser SK, Ndirango A, Sherbondy AJ, Singh R (2011) Commun ACM 54:62

Moore GE (2000) Chapter: cramming more components onto integrated circuits. Morgan Kaufmann Publishers Inc., San Francisco, pp 56–59

Mundy A, Knight J, Stewart T, Furber S (2015) Neural networks (IJCNN), 2015 international joint conference on (2015), IEEE. pp 1–8

Nahmias MA, Shastri BJ, Tait AN, Prucnal PR (2013) IEEE J Sel Top Quantum Electron 19:1–12

Nahmias MA, Tait AN, Shastri BJ, de Lima TF, Prucnal PR (2015) Opt Express 23:26800

Nahmias MA, Tait AN, Tolias L, Chang MP, Ferreira de Lima T, Shastri BJ, Prucnal PR (2016) Appl Phys Lett 108:151106

Nawrocki RA, Voyles RM, Shaheen SE (2016) IEEE Trans Electron Devices 63:3819

Ortín S, Soriano MC, Pesquera L, Brunner D, San-Martín D, Fischer I, Mirasso CR, Gutiérrez JM (2015) Sci Rep 5:14945 EP

Ostojic S (2014) Nat Neurosci 17:594

Paquot Y, Duport F, Smerieri A, Dambre J, Schrauwen B, Haelterman M, Massar S (2012) Sci Rep 2:287 EP

Paugam-Moisy H, Bohte S (2012) Computing with spiking neuron networks. In: Rozenberg G, Bäck T, Kok JN (eds) Handbook of natural computing. Springer, Berlin/Heidelberg, pp 335–376

Perrett DI, Rolls ET, Caan W (1982) Exp Brain Res 47:329

Pfeil T, Grübl A, Jeltsch S, Müller E, Müller P, Petrovici MA, Schmuker M, Brüderle D, Schemmel J, Meier K (2013) Front Neurosci 7:1–17

Pickett MD, Medeiros-Ribeiro G, Williams RS (2013) Nat Mater 12:114

Preston K, Sherwood-Droz N, Levy JS, Lipson M (2011) In: CLEO:2011 laser applications to photonic applications (Optical Society of America), p CThP4

Prucnal PR, Shastri BJ (2017) Neuromorphic photonics. CRC Press/Taylor & Francis Group, Boca Raton

Prucnal PR, Shastri BJ, de Lima TF, Nahmias MA, Tait AN (2016) Adv Opt Photon 8:228

Ramaswami R (1993) IEEE Commun Mag 31:78

Razavi B (2000) Design of analog CMOS integrated circuits. McGraw-Hill Education. New York, NY

Razavi B (2009) IEEE Trans Circuits Syst Regul Pap 56:4

Roelkens G, Liu L, Liang D, Jones R, Fang A, Koch B, Bowers J (2010) Laser Photonics Rev 4:751

Romeira B, Javaloyes J, Ironside CN, Figueiredo JML, Balle S, Piro O (2013) Opt Express 21:20931

Romeira B, Avó R, Figueiredo JL, Barland S, Javaloyes J (2016) Sci Rep 6:19510 EP

Rosenbluth D, Kravtsov K, Fok MP, Prucnal PR (2009) Opt Express 17:22767

Sarpeshkar R (1998) Neural Comput 10:1601

Schemmel J, Briiderle D, Griibl A, Hock M, Meier K, Millner S (2010) In: Proceedings of 2010 I.E. international symposium on circuits and systems (IEEE, 2010), pp 1947–1950

Selmi F, Braive R, Beaudoin G, Sagnes I, Kuszelewicz R, Barbay S (2014) Phys Rev Lett 112:183902

Selmi F, Braive R, Beaudoin G, Sagnes I, Kuszelewicz R, Barbay S (2015) Opt Lett 40:5690

Shainline JM, Buckley SM, Mirin RP, Nam SW (2017) Phys Rev Appl 7:034013

Shastri BJ, Nahmias BJ, Tait AN, Prucnal PR (2014) Opt Quantum Electron 46:1353

Shastri BJ, Nahmias MA, Tait AN, Wu B, Prucnal PR (2015) Opt Express 23:8029

Shastri BJ, Nahmias MA, Tait AN, Rodriguez AW, Wu B, Prucnal PR (2016) Sci Rep 6:19126 EP

Shen Y, Harris NC, Skirlo S, Prabhu M, BaehrJones T, Hochberg M, Sun X, Zhao S, Larochelle H, Englund D, Soljacic M (2017) Nat Photonics. arXiv:1610.02365

Smit M, van der Tol J, Hill M (2012) Laser Photonics Rev 6:1

Snider GS (2007) Nanotechnology 18:365202

Soltani M, Li Q, Yegnanarayanan S, Adibi A (2010) Opt Express 18:19541

Soriano MC, Ortín S, Brunner D, Larger L, Mirasso CR, Fischer I, Pesquera L (2013) Opt Express 21:12

Soriano MC, Brunner D, Escalona-Moran M, Mirasso CR, Fischer I (2015) Front Comput Neurosci 9:1–11

Sorrentino T, Quintero-Quiroz C, Aragoneses A, Torrent MC, Masoller C (2015) Opt Express 23:5571

Spühler GJ, Paschotta R, Fluck R, Braun B, Moser M, Zhang G, Gini E, Keller U (1999) J Opt Soc Am B 16:376

Stewart TC, Eliasmith C (2014) Proc IEEE 102:881

Strogatz SH (2001) Nature 410:268

Sysak M, Liang D, Jones R, Kurczveil G, Piels M, Fiorentino M, Beausoleil R, Bowers J (2011) IEEE J Sel Top Quantum Electron 17:1490

Tait AN, Nahmias MA, Tian Y, Shastri BJ, Prucnal PR (2014a) Photonic Neuromorphic Signal Processing and Computing. In: Naruse M (ed) Nanophotonic information physics. Nano-optics and nanophotonics. Springer, Berlin/Heidelberg, pp 183–222

Tait AN, Nahmias MA, Shastri BJ, Prucnal PR (2014b) J Lightwave Technol 32:3427

Tait AN, Ferreira de Lima T, Nahmias MA, Shastri BJ, Prucnal PR (2016a) Opt Express 24:8895

Tait A, Ferreira de Lima T, Nahmias M, Shastri B, Prucnal P (2016b) IEEE Photon Technol Lett 28:887

Tait AN, Wu AX, de Lima TF, Zhou E, Shastri BJ, Nahmias MA, Prucnal PR (2016c) IEEE J Sel Top Quantum Electron 22:312

Tait AN, de Lima TF, Zhou E, Wu AX, Nahmias MA, Shastri BJ, Prucnal PR (2017) Sci Rep arXiv:1611.02272

Tank D, Hopfield J (1986) IEEE Trans Circuits Syst 33:533

Tapson J, Cohen G, Afshar S, Stiefel K, Buskila Y, Hamilton T, van Schaik A (2013) Front Neurosci 7:153

Taur Y (2002) IBM J Res Dev 46:213

Taur Y, Buchanan D, Chen W, Frank D, Ismail K, Lo SH, Sai-Halasz G, Viswanathan R, Wann HJ, Wind S, Wong HS (1997) Proc IEEE 85:486

The HBP Report (2012) Technical Report (The human brain project)

Thorpe S, Delorme A, Rullen RV (2001) Neural Netw 14:715

Tucker RS (2010) Nat Photonics 4:405

Turconi M, Garbin B, Feyereisen M, Giudici M, Barland S (2013) Phys Rev E 88:022923

Van Vaerenbergh T, Fiers M, Mechet P, Spuesens T, Kumar R, Morthier G, Schrauwen B, Dambre J, Bienstman P (2012) Opt Express 20:20292

Van Vaerenbergh T, Alexander K, Dambre J, Bienstman P (2013) Opt Express 21:28922

Vandoorne K, Dambre J, Verstraeten D, Schrauwen B, Bienstman P (2011) IEEE Trans Neural Netw 22:1469

Vandoorne K, Mechet P, Van Vaerenbergh T, Fiers M, Morthier G, Verstraeten D, Schrauwen B, Dambre J, Bienstman P (2014) Nat Commun 5:3541 EP

Verstraeten D, Schrauwen B, D'Haene M, Stroobandt D (2007) Neural Netw 20:391

Vicsek T (2002) Nature 418:131

Vlasov Y (2012) IEEE Commun Mag 50:s67

von Neumann J (1993) IEEE Ann Hist Comput 15:27

Wieczorek S, Krauskopf B, Lenstra D (1999) Opt Commun 172:279

Wieczorek S, Krauskopf B, Lenstra D (2002) Phys Rev Lett 88:063901

Wieczorek S, Krauskopf B, Simpson TB, Lenstra D (2005) Phys Rep 416:1

Woods D, Naughton TJ (2012) Nat Physics 8:257

Wünsche HJ, Brox O, Radziunas M, Henneberger F (2001) Phys Rev Lett 88:023901

Xiong K, Xiao X, Hu Y, Li Z, Chu T, Yu Y, Yu J (2011) In: Photonics and optolectronics meetings (POEM), vol 8333, pp 83330A–83330A-7

Xu Q, Fattal D, Beausoleil RG (2008) Opt Express 16:4309

Yacomotti AM, Eguia MC, Aliaga J, Martinez OE, Mindlin GB, Lipsich A (1999) Phys Rev Lett 83:292

Yacomotti AM, Monnier P, Raineri F, Bakir BB, Seassal C, Raj R, Levenson JA (2006a) Phys Rev Lett 97:143904

Yacomotti AM, Raineri F, Vecchi G, Monnier P, Raj R, Levenson A, Ben Bakir B, Seassal C, Letartre X, Viktorovitch P, Di Cioccio L, Fedeli JM (2006b) Appl Phys Lett 88:231107

Yamada M (1993) IEEE J Quantum Electron 29:1330

Yu Y, Baltus PG, Van Roermund AH (2011) Integrated 60GHz RF beamforming in CMOS. Springer Science & Business Media. Springer Dordrecht Heidelberg London, New York