Excitable laser processing network node in hybrid silicon: analysis and simulation


Electrical Engineering Dept., Princeton University, Princeton, NJ 08544, USA.

*mnahmias@princeton.edu

Abstract: The combination of ultrafast laser dynamics and dense on-chip multiwavelength networking could potentially address new domains of real-time signal processing that require both speed and complexity. We present a physically realistic optoelectronic simulation model of a circuit for dynamical laser neural networks and verify its behavior. We describe the physics, dynamics, and parasitics of one network node, which includes a bank of filters, a photodetector, and excitable laser. This unconventional circuit exhibits both cascadability and fan-in, critical properties for the large-scale networking of information processors based on laser excitability. In addition, it can be instantiated on a photonic integrated circuit platform and requires no off-chip optical I/O. Our proposed processing system could find use in emerging applications, including cognitive radio and low-latency control.

© 2015 Optical Society of America

OCIS codes: (140.3538) Lasers, pulsed; (200.4700) Optical neural systems; (250.3140) Integrated optoelectronic circuits; (320.7085) Ultrafast information processing.

References and links


#246538

Received 22 Jul 2015; revised 15 Sep 2015; accepted 22 Sep 2015; published 2 Oct 2015

© 2015 OSA


40. S. Le Beux, J. Trajkovic, I. O’Connor, G. Nicoleau, G. Bois, and P. Paulin, “Optical ring network-on-chip...
1. Introduction

Recently, there has been a surge of interest in the hybridization of photonic and electronic physics to achieve unique processing capabilities. In this context, there has been significant research in using laser dynamics for both information processing [1–3] and communication [4,5]. Multiplexing in optical networks have also been exploited for similar gains in performance [6, 7]. A more specific approach involves studying the dynamical property of excitability in lasers, in which discrete pulses are generated in response to perturbations that exceed a threshold [8–11]. Excitable lasers can process information in a way that resembles spiking in biological neuron models. In comparison, however, lasers can exhibit dynamics roughly eight orders of magnitude faster while being mathematically isomorphic to their biological counterparts [12].

In this manuscript, we simulate a unified, realistic model of a reconfigurable excitable laser processor, recently proposed in [13]. This model includes a detailed analysis and simulation of a processing-network node (PNN) as defined in a recent laser networking scheme [14]. Although parts of the signal pathway have been proposed [13–15], no single model has fully characterized the signal pathway of a PNN. The device uniquely allows for a large fan-in (~10s to 100s per unit) without routing or packet switching, and utilizes the ultrafast dynamics of lasers for high bandwidth (~GHz) processing. We describe the photonic circuit techniques used in our approach and simulate the device based on experimentally measured parameters in a standard hybrid III-V/silicon platform [16]. The model allows us to explore critical properties such as energy consumption, cascadability, and signal bandwidth, and verifies that the PNN can be fabricated in a realistic device structure.

Spiking neural networks (SNNs)—systems whose communication channels code information in events rather than bits—have received significant attention as an alternative to the von Neumann paradigm of computation. Hardware SNNs take advantage of distributed, sparse, and robust coding schemes to perform computations to minimize size, weight and power, and have been utilized both as biological network simulators and low-power data processors.

Interconnects in neural network architectures require many-to-one fan-in and multicast communication, as opposed to von Neumann processors, where a number of point-to-point memory-processor communication bottlenecks arise [17]. In distributed systems, processing performance is more closely related to inter-node communication. Distributed electrical wires can introduce performance challenges in a multicast network due to RC and radiative physics, in addition to the typical bandwidth-distance-energy limits of point-to-point electronic links. Since biological-time SNNs typically operate far slower than underlying electronics, many large-scale systems employ a form of time-division multiplexing (TDM) or packet switching, notably, address event representation (AER). TDM and AER abstractions allow networks to have virtual interconnectivity densities that exceed the wire density by a factor related to the sacrificed bandwidth, which can be orders of magnitude [18].

While AER is effective for SNNs running at biological time scales, networks running at faster-than-biological time scales experience new challenges and limitations. The fastest large-scale SNN emulator is capable of an approximately 10 MHz synaptic simulation (10^4 speed-up over biological-time) [19]. In the GHz regime, the gap between this operation bandwidth and feasible real signal bandwidths shrinks, which creates a harsh tradeoff between operation bandwidth and virtual interconnectivity. Neuron emulators relying on electrical wires cannot, in the worst case, simultaneously cope with virtual interconnectivity and GHz signal bandwidths. This is a significant limitation for neuromorphic networks, which rely on high interconnectivity.
densities for complex, parallelized operations. In contrast, it is well known that photonic technology possesses favorable properties for communication and can come to bear on problems in on-chip interconnect performance [20].

The research community has proposed a number of dynamical laser models [12, 21–26]. Unfortunately, many of these models suffer from network or interconnection limitations. Interconnected systems require each node to robustly drive a number of other nodes—referred to as cascadability and fan-in (Fig. 1). If inputs are optically injected into the gain section, the pump wavelength must be shorter than that of the signal for population inversion to be achieved. As a result, network feedback requires wavelength conversion, a fairly expensive operation in terms of size, weight, and power. Alternatively, pumping the intensity directly requires a match with the cavity resonance, limiting inputs and outputs to the same wavelength. Optical fan-in with a single wavelength requires coherent optical summing to avoid beat noise, a practical impossibility in unsynchronized laser systems [27]. These complications would offset many of the advantages of using optical physics over more traditional RF electronic approaches.

To address the issues of cascadability and fan-in for networks of ultrafast spiking laser neurons, a scalable networking architecture was recently proposed [28]. High interconnect density is achieved through the use of wavelength division multiplexing (WDM) signals fanning-in to a single photodetector, which directly drives an electronically-modulated laser neuron (see Fig. 1). In this setting, electronic physics act as a powerful complement to optics, which, as a result of the lack of charge and bosonic properties of photons, are unwieldy for cascadable computation on their own. O/E/O conversion is useful in many contexts, although it is conventionally associated with a high performance cost to demultiplex, digitize, and remodulate signals. In the neural networking context, the electronic link between a photodetector-laser pair can make a cascadable node without necessarily performing these costly functions (see Fig. 2).

This work simulates a unified and realistic model, verifying that fabricating a PNN is within today’s technological capabilities in a hybrid silicon/III-V platform. This is an important stepping-stone towards larger scale simulations and the verification of these concepts in an experimental device. Application of the technology, once developed, would enable a new processing domains, including the manipulation of ultrafast physical phenomena, low-latency control, and—perhaps most pertinently—novel communication and classification schemes in the radio frequency (RF) domain. Examples of the latter include blind source separation, RF fingerprinting, spectral hole exploitation (i.e. non-interfering transmission and reception), and
novel information encoding/retrieval through the direct modulation of carrier waveforms.

2. Methods

As illustrated in Fig. 2, the PNN consists of three primary components: reconfigurable spectral filters, photodetectors, and an excitable laser. Although an inhibitory photodetector is shown in the schematic diagram, the behavior of the devices are similar, aside from a reversal of the ground and signal pads. In this analysis, we only examine the excitatory photodetector pathway for brevity and conciseness. In this scheme, WDM spike signals arrive along a dedicated waveguide. Weights are applied to each channel via a set of tunable spectral filters.

Inputs from other laser neurons are weighted in the optical domain before reaching the photodetector. The photodetectors produce a photocurrent summing the total optical power. Demultiplexing many input channels is not necessary because the incoherent sum of all WDM channels is intentionally computed by the photodetector. The photodetectors receive optical pulses from a network and produce a current signal which modulates the laser carrier injection. The excitable laser performs nonlinear discrimination and regenerates the pulsed signal, analogous to the neural axon hillock. The photodetector front-end proposed here allows for significant signal fan-in, while tunable filters allow adjustments of the weights between neurons, allowing for network reconfigurability.

2.1. Technology Platform

We describe an instantiation of the PNN in the hybrid/III-V platform. This platform includes III-V materials that are bonded to underlying passive silicon photonic interconnection networks. In a typical device, optical modes are hybridized between the silicon and III-V layers simultaneously [29]. Nanostructures necessary for waveguides, resonators, and gratings are fabricated strictly in silicon, while the III-V layers provide optical gain. This platform was chosen for its convenient ability to spatially organize many active elements together onto a single PIC. The analysis below uses realistic parameters derived from experimental papers, and is divided into three primary sections: (a) filtering (i.e. neural weighting) of input WDM-pulsed signals, (b) summation and electrical conversion from photodetector to the adjacent laser, and (c) dynamics present within the laser cavity itself. A schematic of the full modeling structure is shown in Fig. 3.

2.2. Filter Bank

In the proposed WDM network configuration, a large number of processing nodes share a single waveguide and selectively couple light in and out through a bank of adjacent add-drop filters. The pulses coupled from the broadcast waveguide are emitted from various lasers at different wavelengths, and have approximately equal amplitudes. The filter are tuned on and off resonance to control the weight or strength of connection between each element. The result is a series of pulses at different frequencies with an array of different amplitudes [30].

In a given broadcast ring with \( N \) nodes, \( N \) filters are associated with each node to adjust the strength of a total of \( N^2 \) connections. In this analysis, we include only four in a given node for simplicity. We model the transmission function of a resonator drop filter with a Lorentzian function of the form:

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

(1)

where \( \delta \) is the linewidth-normalized frequency, \( Q \) is quality factor (\( Q \approx 10,300 \)), and \( \omega_0 \) is the peak center frequency. Let \( \delta_{\text{run}} \) represent the tunability of each filter, and \( \Delta \delta \) the spacing between adjacent filters. We assume each tunable filter is composed of a single add-drop ring,
Studies of morphology and physiology have pinpointed biological features from polarization switching in a vertical-cavity surface-emitting laser (VCSEL) and from other lasers, each governed by some output power function $P_{\text{out}}(\lambda,t)$ for neuron $j$ at an associated wavelength $\lambda_j$.

There is also a linewidth associated with the output of each laser. Single mode distributed feedback (DFB) lasers, the structures assumed here, tend to have narrow linewidths. Nonetheless, since the output is really a time dependent signal $P_{\text{out}}(\lambda,t)$, the linewidth is fundamentally broadened by modulation, and in addition, can be further broadened by non-idealities (such as carrier-induced chirp). Nonetheless, we will first consider an ideal case in which bandwidth-limited pulses are assumed. This allows us to approximate the output power spectrum as $S_{\text{out}[i]}(\lambda,t) = P_{\text{out}[i]}(t)\delta(\lambda - \lambda_i)$ for each laser $i$ in a given loop where $\delta(\lambda)$ is a Dirac delta function centered at $\lambda_i$. Then the broadcast loop signal $B(\lambda,t)$ contains a sum of all output power spectra, i.e.

$$B(\lambda,t) = \sum_i S_{\text{out}[i]}(\lambda,t)$$  \hspace{1cm} (2)

Each filter $i$ in a given filter bank is associated with a wavelength $\lambda_i$ corresponding to a laser $i$ within the broadcast space. Although there will be some crosstalk due to the Lorentzian shape of the filter, we have minimized crosstalk by design ($< 13$ dB), so we can neglect its effect to first order. The power spectrum of the signal dropped from the broadcast waveguide for neuron...
$j$ will take the form:

$$S_{in[j]}(\lambda, t) = \sum_i T_{ij}(\lambda) S_{out[i]}(\lambda, t)$$  \hspace{1cm} (3)$$

where $T_{ij}(\lambda)$ is the transmission function of ring $i$ for neuron $j$. Therefore, for a given photodetector spectral responsivity linear system $R_{\lambda}$, the current produced by the photodetector is as follows:

$$i_{p[j]}(t) = \sum_i \int \lambda R_{\lambda} \{ T_{ij}(\lambda) S_{out[i]}(\lambda, t) \} d\lambda$$  \hspace{1cm} (4)$$

Based on our assumptions, we can approximate the filter linewidth as significantly larger than the signal bandwidth. The 40 ps sech$^2$ pulses used in this simulation have transform-limited bandwidth of $\Delta \delta = .42$ (0.63 nm), which is much smaller than the bandwidth of the filters (as shown in Fig. 4(b)), allowing this approximation to be valid. We assume the photodetector response is approximately spectrally flat over the C-band [31], and neglect carrier diffusion limitations to frequency response, which are small compared to the RC limitations included in this model. This allows us to approximate the responsivity as a constant $R_{\lambda} \{ P(t) \} = R_{PD} P(t)$. Therefore, we can simplify the photodiode current to the following expression:

$$i_{p[j]}(t) \approx \sum_i R_{PD} T_{ij}(\lambda_i) P_{out[i]}(t)$$  \hspace{1cm} (5)$$

A simulation of a filter bank with four filters is shown in Fig. 4. Pulses of different frequencies experience different weights—or amplitude shifts—based on the spectral position of each tunable ring. Based on the parameters used here, there can be a total of $N = 34$ filters per node within 45 nm gain band in hybrid III-V/Si. This number is synonymous with the upper bound on the fan-in per neuron. The networks themselves can be designed to have much larger numbers of neurons by using multi-loop strategies that can sidestep fan-in limitations [28]. In addition, the upper limit can be expanded: 62 channels have been demonstrated simultaneously in an experimental silicon photonics platform, a number that can be extended further using more advanced techniques [32]. Nonetheless, fan-in puts an upper limit on the complexity of the computation that a single unit can perform. The scalability of neural fan-in must be balanced with respect to spectral stability of the lasers, the linewidth of the pulses and the Q factor of the filters (this point is expanded in the Discussion section).

### 2.3. Electronic Junction

The analysis of the electrical junction parasitics describes the signal as it is received from the photodetector and drives the adjacent excitable laser. It is converted from a series of pulses at different wavelengths to a single current signal (Fig. 4). Although current diffusion, saturation, and other nonidealities can also play a role in the limiting the frequency response of the device, RC charging times tend to dominate [31]. We model the electronic junction as the lumped circuit (Fig. 3) using parasitic values that were determined in several experimental papers [29, 33, 34].

In this analysis, we only consider the bandwidth of a single photodetector driving a laser for simplicity, although multiple photodetectors may also be used for push-pull excitatory and inhibitory inputs, if desired. Multiple pulses of different wavelengths are incident on the photodetector, which provides the role of summation and converts signals to current pulses. A schematic of the junction is shown in Fig. 2, including a superimposed image of the parasitic circuit model. The purpose of the model is to find the relationship between the photodetector current $i_p(t)$ and laser driving current $i_e(t)$:

$$i_e(t) = F \{ i_p(t) \}$$  \hspace{1cm} (6)$$
A great variety of possible weight analogous to a neural weight. The filters of a given receiver simultaneously and with intermediate strengths between 0% and 100% of the local states of filters and not a state of the transmission network, a WDM protocol called broadcast-and-select is one way to efficiently use the full capacity of a wavelength division multiplexed channelization of the spectrum. The ability to control each connection, each weight, in an amplitude and asynchronous in time. Spike timing is the linear transfer function of the modeled circuit. These include photonic crystal and nano-beam lasers, distributed bragg reflector (DBR) lasers, and vertical cavity surface emitting lasers (VCSELs). This archetypal model is called a general interaction, thresholding). The contrasting all-to-all nature of the network interconnectivity pattern, so the role of neurons is predominantly computational (weighted addition, integration, thresholding). The channel subsets that are dropped are determined thermally or electronically. While the general expression for the complex impedance of the link is more involved, in practice, the dominant parasitics are the capacitance of the metal wire $C_M$ and the contact resistances $R_{PD}, R_{L}$. The metal wire capacitance $C_M$ is the easiest to adjust lithographically, either by changing the height of the oxide layer or the area occupied by the metal bridge to change the characteristics of the junction.

A common first order model for synaptic dynamics is an integrator with decay [35]:

$$\frac{ds}{dt} = -\frac{s}{\tau} + I(t)$$

where $I(t)$ is the input, $\tau$ is the time constant and $s(t)$ is the synaptic variable. A neuron will typically sum signals from multiple synapses $s_i(t)$ and receive them as inputs. In biological systems, synaptic variables typically represent neurotransmitter concentrations. In our case, it represents the RC charged signal. This behavior takes place independently of temporal integra-

$$\lambda_i : \text{P}_{\text{in}[i]}$$

$$\lambda_j : \text{P}_{\text{out}[j]}$$

$$\lambda_k : \text{P}_{\text{out}[k]}$$

$$\lambda_l : \text{P}_{\text{out}[l]}$$

where $F$ is the linear transfer function of the modeled circuit.

A simulation of multiple input pulses is shown in Fig 5. The laser is pumped with a stable current source $I_p$, while the PD is reverse biased with a large voltage (>5 V) to offset the influence of $I_p$. Responsivity is assumed to be 0.81A/W [16]. The junction is kept fairly short (~100s of $\mu$m) to avoid transmission line effects.

While the general expression for the complex impedance of the link is more involved, in practice, the dominant parasitics are the capacitance of the metal wire $C_M$ and the contact resistances $R_{PD}, R_{L}$. The metal wire capacitance $C_M$ is the easiest to adjust lithographically, either by changing the height of the oxide layer or the area occupied by the metal bridge to change the characteristics of the junction.

A common first order model for synaptic dynamics is an integrator with decay [35]:

$$\frac{ds}{dt} = -\frac{s}{\tau} + I(t)$$

where $I(t)$ is the input, $\tau$ is the time constant and $s(t)$ is the synaptic variable. A neuron will typically sum signals from multiple synapses $s_i(t)$ and receive them as inputs. In biological systems, synaptic variables typically represent neurotransmitter concentrations. In our case, it represents the RC charged signal. This behavior takes place independently of temporal integra-
Table 1. Hybrid/III-V Laser Parameters

<table>
<thead>
<tr>
<th>Param.</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_g$</td>
<td>gain section volume</td>
<td>$1.68 \times 10^{-11} \text{ cm}^3$</td>
</tr>
<tr>
<td>$V_{\alpha}$</td>
<td>SA section volume</td>
<td>$3.36 \times 10^{-12} \text{ cm}^3$</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>QW conf. factor</td>
<td>0.056</td>
</tr>
<tr>
<td>$n_0$</td>
<td>transparency carrier density</td>
<td>$1.75 \times 10^{18} \text{ cm}^{-3}$</td>
</tr>
<tr>
<td>$v_g$</td>
<td>group velocity</td>
<td>$c/3.49$</td>
</tr>
<tr>
<td>$g_0$</td>
<td>QW gain coefficient</td>
<td>$966 \text{ cm}^{-1}$</td>
</tr>
<tr>
<td>$\tau_g$</td>
<td>gain carrier lifetime</td>
<td>1.1 ns</td>
</tr>
<tr>
<td>$\tau_{\alpha}$</td>
<td>SA carrier lifetime</td>
<td>100 ps</td>
</tr>
<tr>
<td>$\tau_{ph}$</td>
<td>photon lifetime</td>
<td>2 ps</td>
</tr>
<tr>
<td>$n_{sp}$</td>
<td>spontaneous noise factor</td>
<td>2</td>
</tr>
<tr>
<td>$I_g$</td>
<td>gain pumping current</td>
<td>21 mA</td>
</tr>
<tr>
<td>$\eta_i$</td>
<td>injection efficiency</td>
<td>0.6</td>
</tr>
<tr>
<td>$\eta_c$</td>
<td>differential quantum efficiency</td>
<td>0.26</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>lasing wavelength</td>
<td>1550 nm</td>
</tr>
</tbody>
</table>

Information in the soma. The synaptic time constant thus has a correspondence with this circuit’s RC time constant.

As shown in Fig. 4, the electronic junction does not significantly degrade the bandwidth of input pulses with FWHMs on the order of $\sim$ps. In fact, the RC time constant can be as low as 30 ps using standard lithographic techniques [15]. This shows that optical injection is not necessary to maintain the information content in high bandwidth pulses, if the electrical junction is short. Nonetheless, a slower synaptic time constant—which is useful for various processing applications—is still possible by engineering a larger capacitance across the wire junction.

2.4. Laser Neuron

2.4.1. Dynamics

The dynamical system underlying the behavior of our processing model is a gain-absorber cavity, describing single mode lasers with gain and saturable absorber (SA) sections. Despite its simplicity, it can exhibit a large range of possible behaviors [9], and has been investigated in various contexts as the basis for an optical processor [22]. The system in its simplest form can be described using the following undimensionalized equations:

\[
\begin{align*}
\dot{G}(t) &= \gamma_G[A - G(t) - G(t)I(t)] + \theta(t) \\
\dot{Q}(t) &= \gamma_Q[B - Q(t) - aQ(t)I(t)] \\
\dot{I}(t) &= \gamma_I[G(t) - Q(t) - 1]I(t) + \varepsilon f(G)
\end{align*}
\]

where $G(t)$ models the gain, $Q(t)$ the absorption, and $I(t)$ the laser intensity. $A$ is the gain bias current, $B$ is the absorption level, $\gamma_G$ is the gain relaxation rate, $\gamma_Q$ is the absorber relaxation rate, $\gamma_I$ is the inverse photon lifetime, and $a$ is a differential absorption relative to the gain factor. We represent spontaneous noise via $\varepsilon f(G)$, for small $\varepsilon$, and time-dependent input perturbations as $\theta(t)$. 

#246538

Received 22 Jul 2015; revised 15 Sep 2015; accepted 22 Sep 2015; published 2 Oct 2015
© 2015 OSA
Under a desired parameter regime, the internal dynamics can be compressed so that pulse generation is instantaneous. The behavior simplifies to [12]:

\[
d\frac{dG(t)}{dt} = -\gamma_e(G(t) - A) + \theta(t);
\]

if \( G(t) > G_{\text{thresh}} \), then release a pulse, and set \( G(t) \rightarrow G_{\text{reset}} \)

where the input \( \theta(t) \) can include spike inputs of the form \( \theta(t) = \sum \delta(t - \tau_i) \) for spike firing times \( \tau_i \), \( G_{\text{thresh}} \) is the gain threshold, and \( G_{\text{reset}} \sim 0 \) represents the gain at transparency. This system is analogous to a leaky integrate-and-fire (LIF) neuron model, commonly employed in computational neuroscience for modeling biological neural networks. Although it is one of the simpler spike-based models, the LIF model is capable of universal computations [36], and is capable of coding information in the timing between spikes [37].

### 2.4.2. Device

We consider a two-section distributed feedback (DFB) excitable laser in the hybrid platform. We use a DFB cavity to guarantee a single longitudinal lasing mode, defined through the lithographic definition of lasing wavelength via silicon-on-insulator grating pitch. The quarter wavelength-shifted defect is placed asymmetrically in the cavity to direct the light in one direction. An ion implantation region electrically isolates the semiconductor absorber and gain sections, and a separate implant provides a smaller lifetime in the absorber section. Small modulation currents that exceed the \( Q \)-switch threshold will trigger large pulse discharges (see Fig. 5(a)).

We define the effective gain and absorption factors for each section as:

\[
\tilde{g}(n_{g[j]}) = \Gamma \frac{V_g g_0}{n_0} (n_{g[j]} - n_0)
\]

\[
\tilde{\alpha}(n_{\alpha[j]}) = \Gamma \frac{V_g g_0}{n_0} (n_0 - n_{\alpha[j]})
\]

where \( \Gamma \) denotes the confinement factor of the QW sections with the transverse mode, \( V_g \) is the waveguide group velocity, \( g_0 \) is the gain coefficient of the material, \( n_0 \) is the transparency density and \( n_g, n_\alpha \) denote the carrier concentrations in the gain and absorber sections, respectively.

For laser \( j \), we can describe the internal cavity dynamics using the following system of equations:

\[
\frac{dN_{ph[j]}}{dt} = \left[ \tilde{g}(n_{g[j]}) - \tilde{\alpha}(n_{\alpha[j]}) - \frac{1}{\tau_{ph[j]}} \right] N_{ph[j]} + \tilde{g}(n_{g[j]}) \frac{n_e}{V_g} \tag{10a}
\]

\[
\frac{dn_{g[j]}}{dt} = \frac{I_{g[j]} + i_{c[j]}(t)}{eV_g} - \frac{n_{g[j]}}{\tau_g} - \tilde{g}(n_{g[j]}) \frac{N_{ph[j]}}{V_g} \tag{10b}
\]

\[
\frac{dn_{\alpha[j]}}{dt} = -\frac{n_{\alpha[j]}}{\tau_\alpha} + \tilde{\alpha}(n_{\alpha[j]}) \frac{N_{ph[j]}}{V_\alpha} \tag{10c}
\]

where \( N_{ph}(t) \) is the total number of photons in the cavity, \( n(t) \) is the number of carriers, \( n_0 \) is the transparency carrier density, \( V \) is the cavity volume, \( \Gamma \) is the confinement factor, \( \tau \) is the carrier lifetime, \( \tau_{ph} \) is the photon lifetime, \( n_e \), the spontaneous emission factor, and \( i_{c}(t) \) represents the electrical modulation in the gain provided by the photodetector system, separate from the constant current bias \( I_g \). Subscripts \( g \) and \( \alpha \) identify the active and absorber regions. Parameter values are shown in Table 1 and based on those found in [31, 34]. The absorber is
assumed to have a faster lifetime, which helps improve the dynamics and consistency of the input signals [13]. The equations can be simplified to the undimensionalized set of equations through a set of variable substitutions [12, 38]. The results are shown in Fig. 5.

The output power of laser \( j \) can be computed via:

\[
P_{\text{out}[j]}(t) = \eta_c \frac{hc}{\tau_{ph}} N_{\text{ph}[j]}(t)
\]

for output coupling efficiency \( \eta_c \), photon lifetime \( \tau_{ph} \), Planck’s constant \( h \), speed of light \( c \), wavelength \( \lambda_j \), and photon number \( N_{\text{ph}}(t) \). The resulting single-mode power spectrum of the laser takes the form \( S_{\text{out}[j]}(\lambda, t) = P_{\text{out}[j]}(t) \delta(\lambda - \lambda_i) \), as mentioned in section 2.2. The signal \( S_{\text{out}[j]} \) is coupled into the broadcast loop, becoming the input of other PNNs.

2.5. Non-Idealities

Although this analysis has considered the physics of the signal pathway to first order, there are many second order non-idealities that will require further study and consideration. Discussed here is a brief overview of the major sources that could lead to decreased performance. First, let us consider fan-in limitations. In addition to the fundamental transform-limited bandwidth \( \delta\lambda_B \), there is also temperature-induced linewidth fluctuations \( \delta\lambda_T \) and carrier-induced laser chirp \( \delta\lambda_C \). If the filters have a linewidth of \( \delta\lambda_F \), we must space filters by at least

\[
(\Delta\lambda)^2 \approx (\delta\lambda_B)^2 + (\delta\lambda_T)^2 + (\delta\lambda_C)^2 + (\delta\lambda_F)^2
\]

(12)

to avoid significant interference. Closely spaced, overlapping filters can unintentionally allow the same signal to interfere with itself, resulting in coherent crosstalk. This manifests as phase differences \( \delta\phi \) between each path becoming visible as amplitude noise \( \delta P \) with an SNR that depends on the magnitude difference between the filtered signals.

Therefore, for a given a channel spacing \( \Delta\lambda \) and gain bandwidth \( G_\lambda \) (typically around \( \sim 50 \text{ nm} \)), the maximum channel fan-in capacity \( C_f \) of the system can be defined as:

\[
C_f \approx \frac{G_\lambda}{\delta\lambda}
\]

(13)

This puts a fairly hard limit on the fan-in of neurons. In the very best case, (i.e. bandwidth-limited channels with 10 ps pulses in a 50 nm gain bandwidth), we can expect a maximum of \( \sim 200 \) channels. Although it is a hard limit, this number is still very high compared to what is possible in electronics, and still unmatchable at such enormous signal bandwidths (i.e. \( \sim 40 \text{ GHz} \)). In addition, networks using clever organizational strategies can include many more PNNs than the fan-in limit, as mentioned earlier in the manuscript.

We must also consider noise sources that can compromise the SNR of the signal as it travels through the system, which has not been studied here. There are many sources of noise to consider including amplified spontaneous emission (ASE) noise, thermal noise, and shot noise for both the optical and electrical signal. Since the laser pump current \( I_p \) is the largest signal in the pathway, we can expect that its shot noise contribution \( \sigma \) will dominate as input power is reduced. Nonetheless, all noise sources can play a contribution in reducing the signal SNR as it travels through the pathway and obfuscates the lasers ability to make a discrete decision based on its input. The effect of noise on this pathway remains to be studied, although it will not be significantly higher than what is already seen in directly modulated laser diodes and photodetectors.
3. Discussion

The model developed in this paper—including the filter bank, electrical junction, and laser neuron—enables the exploration of both analog signal properties and the effect of physical design decisions on the behavior of networked PNNs. An example simulation of the complete model with WDM optical-in and optical-out is shown in Figs. 4 and 5. The front-end weight bank and photodetector in Figs. 4(a)–4(c) act to generate an electronic representation of the weighted sum of WDM inputs carrying FWHM = 40 ps pulses. This signal that modulates the laser neuron is shown in Fig. 4(d), after traversing the parasitic circuit model from Fig. 3. While some pulse spreading is visible (FWHM = 56 ps), pulse amplitude and timing information is clearly maintained when realistic parasitic values are used. Figure 5 illustrates the internal dy-
Fig. 5. Simulated internal dynamics of the laser neuron in response to the modulation signal in Fig 4.d. (a) input electrical modulation $i_e(t)$ (black line) causes the release of a large optical pulse near 6ns (red line). Inset shows the magnified output spike, which is approximately a $sech^2$ pulse with FWHM=16.5 ps. (b) Simulated carrier densities in both the gain (blue line) and saturable absorption (purple line) sections of the laser. The dashed line represents the transparency carrier density, $n_0$.

Dynamics of the excitable laser neuron in response to this modulation signal. Small perturbations to the gain population produce no optical output, but a larger input causes saturation of the SA and release of a short pulse (FWHM=16.5 ps), even narrower than the original input pulses. This is a testament to the system’s regenerative properties, which can continue to propagate pulses without their eventual degradation [12].

This simulation indicates that the PNN design meets a number of important scalability criteria: the dynamical processing model is both cascadable and restorative, the signal pathway can accept large fan-in and perform summation, elements are compatible with a high-volume multiplexing scheme, and the processing nodes are constructed from standard elements that can be implemented on PICs. Prior work on spiking laser neurons has largely focused on single laser dynamical phenomena, while suggesting the possibly of all-optical processing networks. No realistic approach for many-to-one interconnection with optically injected lasers has been proposed, and the well-known challenges of all-optical modulation (wavelength conversion, phase noise accumulation, two-beam beat interference) remain a question in these systems.

The electronic mediation in the PNN enables a cascadable node based on a spiking laser neuron, but does not bring the typical downsides of O/E/O conversion in digital communication systems because transduction is separated from receiving (i.e. sampling, quantization, retiming). The energy and cost of electronic conversion in optical systems results largely from high-speed clocked transistor circuitry [39] and the need to demultiplex WDM channels before conversion [40]. In our system, electronic transduction does not regenerate or receive signals, but instead exploits the electronic physics for intermediate analog processing. The conversion between optical and electronic domains curtails the propagation of optical phase noise and the need for direct wavelength conversion, eliminating two major barriers facing scalable optical...
computing [41]. At the same time, the temporary conversion to the electrical domain does not significantly degrade the signal, as this work has simulated. Every device in the primary signal pathway performs both physical and computational roles, resulting in a robust, ultrafast, and efficient signal pathway.

Utilization complementary physics along the signal pathway is not unlike this pathway in biological neurons, in which electrical action potentials are converted to chemical signals called neurotransmitters upon reaching a synapse (the junction between neurons). Chemical signaling is a relatively short-range process, but it introduces much more functionality compared to direct electrical modulation (e.g. both excitation and inhibition, a variety of synaptic time constants, and the ability to adapt the connection strength). The post-synaptic neuron body passively accumulates the effect of many synapses, and its internal spiking dynamics generate a single channel of electrical action potentials, which is suitable for long-distance transmission. This process is roughly paralleled by O/E/O conversion in the PNN: the balanced photodetector and analog electrical link enable cascadable summation of multiwavelength inputs, from which an excitable laser generates a new spiking optical signal. This functionality does not impact bandwidth performance because the electrical link can be very short-range, similar to how the neurotransmitter link is very short in the biological pathway.

4. Conclusion

We have investigated a device that can act as a node in an ultrafast photonic neural network, using a computational model that fully exploits the temporal resolution of optical signals. The architecture is physically enabled by three primary discoveries. First, a laser with gain and saturable absorption sections behave analogously to a biological neuron, but with dynamics and temporal resolutions on the order of nanoseconds and picoseconds, respectively. Secondly, a WDM broadcast-and-weight networking approach allows a large number of signals to be superimposed on a single waveguide where connection strengths are easily configured by tunable filter banks. Finally, a short receiver-less electronic link can carry a WDM fan-in signal without significant distortion or loss. By using the full model constructed here that takes into account many physical effects, we show that these optical subcircuits constitute a complete PNN capable of participating in a cascadable, scalability photonic neural network.

Although the PNN promises several unique properties advantageous for processing, there still remain aspects that require exploration as the number of elements increase—these include the reliability and configurability of WDM filter-based weighting, the effect of noise accumulation, the effect of transform-limited pulses and chirp on the interconnection network density, and perhaps most importantly, power consumption and the effect of temperature on a large number of active elements clustered together on a PIC. Nonetheless, this represents an important direction in the consideration of scalability for photonic neural networks. Further developments could pave the way for processing systems that could perform real-time, complex processing of ultrafast (GHz) signals in real-time, which has the potential to revolutionize radio cognition or the control of ultrafast physical phenomena.

Funding Information

The authors acknowledge the support of the National Science Foundation Graduate Research Fellowship Program (NSF GRFP) and the Banfing Postdoctoral Fellowship administered through the Natural Sciences and Engineering Research Council of Canada.