A TeraMAC Neuromorphic Photonic Processor

Mitchell A. Nahmias, Hsuan-Tung Peng, Thomas Ferreira de Lima, Chaoran Huang, Alexander N. Tait, Bhavin J. Shastri and Paul R. Prucnal
Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544 USA, email: mnahmias@princeton.edu

***Abstract*—We show that an integrated laser neuron can exhibit extraordinary low latency ($<1\,\mathrm{ns}$) and speed ($\sim 1 \times 10^{12}\,\mathrm{MACs/s}$ per device) compared to state-of-the-art processors in digital electronics. We experimentally demonstrate positive (excitatory) and negative (inhibitory) inputs with $8\,\mathrm{x}$ wavelength channels, and efficiency ($<1\,\mathrm{pJ/MAC}$) during closed-loop operation.**

I. Introduction

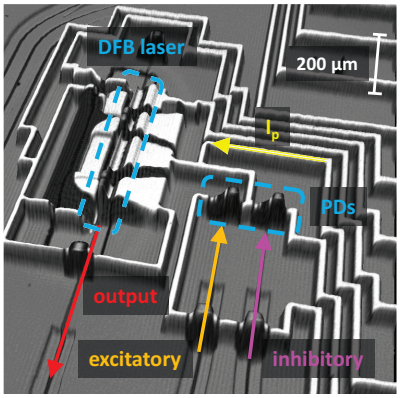
Recent demand for deep learning hardware has lead to immense communication and computational requirements. Specialized processors, such as the Tensor Processing Unit (TPU), are overtaking generalized computing processors such as CPUs or GPUs in popularity and performance [1]. A key metric in such hardware is energy consumption, which includes both data movement and computation, the latter of which consist largely of matrix computations. The constituent components of matrix computations are multiply-accumulate (MAC) operations, i.e., operations of the form $a = a + w \times x$. Photonics has the potential to address both the communication and computation bottlenecks: (1) optical interconnects can provide low-energy communication channels unconstrained by distance [2], while (2) photonic matrix operations scale favorably, since energy consumption is proportional to the number of channels rather than the number of MACs [3].

Here, we test and explore the properties of a laser neuron processor, instantiated on a indium phosphide photonic integrated circuit (PIC) platform. This laser neuron exhibits the biologically relevant property of *spiking* at a far faster time scale than biological ($> 10^8$) or electrical ($> 10^3$) systems. In addition, it exhibits extraordinarily performance compared to the state-of-the-art in deep learning and neuromorphic electronic hardware: a single device can process $\sim 1 \times 10^{12}\,\mathrm{MACs/s}$ per second, with an energy efficiency of $\sim 270\,\mathrm{fJ}$ per MAC operation.

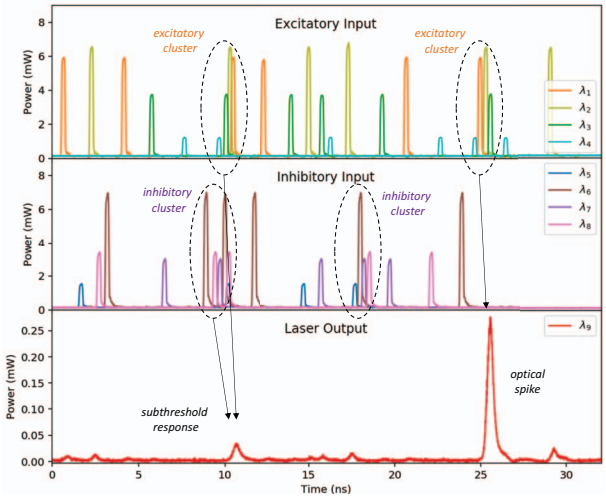
II. Multi-Wavelength Functionality

The processor is designed to be compatible with a wavelength network protocol called Broadcast-and-Weight (B&W) [4]. This protocol utilizes wavelength division multiplexing (WDM) to allow for scalable interconnects between laser neuron processors. Large networks (> 100) can be instantiated on-chip, requiring only a small number of waveguides without any waveguide crossings [5]. We designed a composite device structure, fabricated in a standard indium phosphide PIC platform by the Heinrich Hertz Institute. The structure is shown in Fig. 1a.

Our experimental demonstration used a total of $8\,\mathrm{x}$ wavelength channels with independent spiking signals: $4\,\mathrm{x}$ inputs were associated with each photodetector. The laser is operated just below the lasing threshold s.t. the loss exceeds the overall gain. An inhibitory and excitatory photodetector provide negative and positive input via a push-pull configuration. Only when a large cluster of excitatory pulses arrive closely spaced in time with no inhibition does the laser release an optical pulse ($\sim 300\,\mathrm{ps}$ width). The presence of negative feedback via



(a)



(b)

Fig. 1: (a) Topographic micrograph of an integrated laser neuron. Excitatory and inhibitory inputs are incident on a balanced photodetector (PD) pair, which drives a distributed feedback (DFB) laser biased with current I_p. (b) Device tested with $8\,\mathrm{x}$ independent wavelength channels $\lambda_i \ldots \lambda_8$. All wavelengths are in the telecommunications C-band.

978-1-5386-5358-6/18/$31.00 ©2018 IEEE

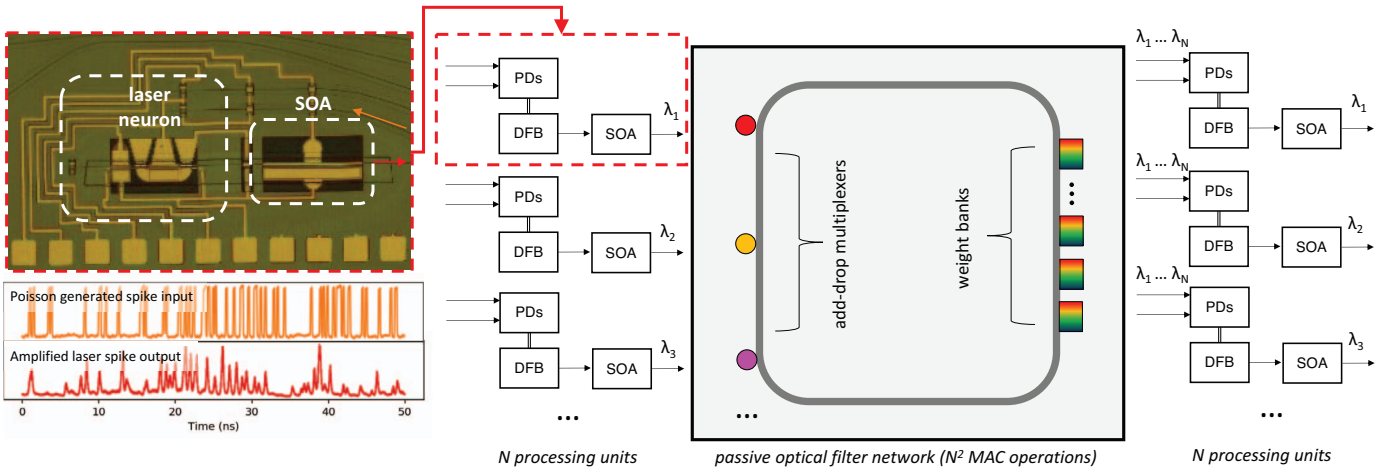


Fig. 2: Left: micrograph of a laser neuron with a semiconductor optical amplifier. Left bottom: measured input trace, generated using a Poisson process, and measured output during closed-loop gain. Right: schematic of a hypothetical system with an on-chip passive optical network. N^2 MAC operations are performed, but energy cons. scales with the processor number N.

an inhibitory cluster prevents a pulse from being released. This basic functionality, shown in Fig. 1b, emulates a Leaky Integrate-and-Fire neuron, a Turing-complete spiking model well known in the field of computational neuroscience [6].

The system can scale beyond the tested $8\,\mathrm{x}$ channels to encompass the full >100 channels seen in dense WDM. Recent work in the developming of microring filters has shown that up to $\sim$200 channels are possible [7]. If the laser devices are interfaced with passive filters using interposer technology or a hybrid silicon/III-V platform, we can scale the system with many integrated components. With $N\sim200$ and a refractory period of $\sim0.2\,\mathrm{ns}$, a single laser neuron can perform approximately $\sim1\times10^{12}\,\mathrm{MACs/s}$.

III. Energy Consumption

For a desired *mean precision* per channel, the power consumption scales with the number of channels N, rather than the number of MAC operations N^2. This is because passive circuits do not in principle consume energy. As shown in Fig. 2, a series of N laser neurons input signals into a passive network, which redistributes those signals (i.e., multiplies them by a weight matrix) before the next stage. As long as each set of N processors can amplify their output and compensate for passive losses—typically $<3\,\mathrm{dB}$ on-chip—the system can continue processing information regeneratively.

Cascadable operation requires meeting the closed-loop gain condition: the output power must exceed the input. To achieve this condition, we designed a laser neuron with a semiconductor optical amplifier (SOA) affixed to the output of the laser, a micrograph of which is shown in Fig. 2. To measure the required energy needed to pump a SOA, we generated input pulses via a Poisson process to the excitatory PD with pulse widths of $0.2\,\mathrm{ns}$ and a Poisson rate of $\lambda_p = 1\,\mathrm{GHz}$. We adjusted the SOA current to meet the closed-loop gain condition. which occurred at $I_{\mathrm{SOA}} = 105\,\mathrm{mA}$ at a voltage of $2.50\,\mathrm{V}$. Since the laser and photodetector together consumed $<10\,\mathrm{mW}$ of power, the SOA energy consumption dominates. Nonetheless, with a TeraMAC processing capacity, this amounts to $\sim270\,\mathrm{fJ}$ per MAC. Another key advantage is that since the signals are encoded on light by default, no additional energy is consumed moving data from one place to another—this is typically a major source of power loss in electronics.

IV. Conclusion

We have fabricated and tested a laser neuron, instantiated in a photonic integrated circuit platform. It displays many favorable properties that sidestep many of the limitations inherent in digital electronics. Note that—as a proof-of-concept—there are many ways in which the performance can be improved. For example, stronger absorption dynamics can lead to shorter pulses; microwave loss between the photodetector to laser can be reduced by increasing the impedance of the input line; smaller devices can lead to higher conversion efficiency, and the use of novel materials such as graphene could lead to much faster ($>100\,\mathrm{TMACs/s}$) operation [8]. Further optimizations may eventually make the amplifier unnecessary, paving the way for groundbreakiong energy efficiencies ($<1\,\mathrm{fJ/MAC}$) and performance for next generation deep learning systems.

References

[1] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings*. ACM, 2017, pp. 1–12.

[2] D. A. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *JLT*, vol. 35, no. 3, pp. 346–396, 2017.

[3] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, pp. 441, 2017.

[4] A. N. Tait *et al.*, "Broadcast and weight: an integrated network for scalable photonic spike processing," *JLT*, vol. 32, no. 21, pp. 3427–3439, 2014.

[5] P. R. Prucnal *et al.*, *Neuromorphic photonics*. CRC Press, 2017.

[6] M. A. Nahmias *et al.*, "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," *IEEE JSTQE*, vol. 19, no. 5, 2013.

[7] A. N. Tait *et al.*, "Two-pole microring weight banks," *Optics Letters* (accepted), 2018.

[8] B. J. Shastri *et al.*, "Spike processing with a graphene excitable laser," *Scientific Reports*, vol. 6, pp. 19 126, 2016.