SILICON PHOTONICS FOR ARTIFICIAL INTELLIGENCE APPLICATIONS

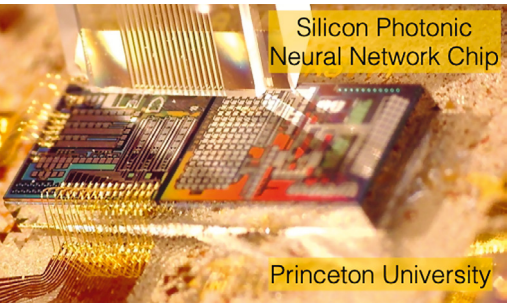
Bicky A. MARQUEZ[1,*], Matthew J. FILIPOVICH[1], Emma R. HOWARD[1], Viraj BANGARI[1], Zhimu GUO[1], Hugh D. MORISON[1], Thomas FERREIRA DE LIMA[2], Alexander N. TAIT[2,3], Paul R. PRUCNAL[2], Bhavin J. SHASTRI[1,2]

[1] Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston, ON KL7 3N6, Canada

[2] Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

[3] Applied Physics Division, National Institute of Standards and Technology, Boulder, CO 80305, USA

***bama@queensu.ca**



Artificial intelligence enabled by neural networks has enabled applications in many fields (*e.g.* medicine, finance, autonomous vehicles). Software implementations of neural networks on conventional computers are limited in speed and energy efficiency. Neuromorphic engineering aims to build processors in which hardware mimic neurons and synapses in brain for distributed and parallel processing. Neuromorphic engineering enabled by silicon photonics can offer sub-nanosecond latencies, and can extend the domain of artificial intelligence applications to high-performance computing and ultrafast learning. We discuss current progress and challenges on these demonstrations to scale to practical systems for training and inference.

https://doi.org/10.1051/photon/202010440

Analog computing has recently been considered a potential avenue to decrease energy and time requirements for executing algorithms such as deep neural networks. Analog special-purpose hardware requires the manufacturing of machines to physically model each individual component of such networks. This proves to be a significant challenge as current deep networks scale up to thousands or even billions of neurons to solve complex artificial intelligence (AI) tasks. To enable the use of analog machines to map brain circuitry, the functions of biological neurons must be modelled. The most common neural models are spiking artificial neurons and perceptrons. While spiking artificial neurons are biologically realistic, the field of AI is currently perceptron-based [1].

Perceptrons implement multiply-accumulate (MAC) operations. MAC operations serve to quantify the number of multiplications and additions required to run deep networks. A perceptron of M inputs can perform M MAC operations per time step. Multiple MAC operations can be executed in parallel to implement any type of artificial neural network (ANN). MACs are currently the most burdensome hardware bottleneck in ANNs; for instance, the deep network AlexNet requires 724 million MACs to solve ImageNet [2].

The photonic platform is currently one of the most promising technologies to tackle the expensive calculations performed by deep networks. Silicon photonics offers high-scalability, high-bandwidth, low-footprint, and low-energy consumption [3]. The high-bandwidth and multiwavelength parallel properties of light allow for optical information processing at a high data rate. The ability of neuromorphic photonic systems to provide substantial improvement in our computing capabilities is moving ever closer, with, potentially, PetaMac/second/mm^2 processing speeds.

In this article, we describe a photonic scheme that can perform parallel MAC operations on-chip and introduce two photonic platforms that allow for

AI hardware acceleration: i) a special-purpose photonic architecture for executing the direct feedback alignment (DFA) algorithm for neural network training [6], and ii) an implementation of a Long-Short Term Memory (LSTM) neural network [7]. Both proposed designs offer fundamental speed and bandwidth advantages over digital electronic implementations.

BACKGROUND: NEUROSCIENCE AND COMPUTATION

Digital computers are typically computing systems that perform logical and mathematical operations with high accuracy. Nowadays, such complex systems significantly outweigh human capabilities for calculation and memory. Nevertheless, if we were to compare a human agent with a digital machine, we would see that there are many abstractions that should be made to perform one-to-one comparisons. Such abstractions assume that human cognitive processes are completely procedural and follow standard logic. However, most human cognitive acts do not follow a set of well-defined instructions. Therefore, a one-to-one mapping between human and digital computers might not be suitable.

Analog neuromorphic computing approaches might be more suited to mimic human brain processes. The goal is to create a one-to-one mapping between the neural system and the analog machine, where each biological quantity is modelled by an equivalent analog artificial model. For an architecture such as the human brain, this could be a demanding requirement. The human brain contains approximately 100 billion neurons and 100 trillion synaptic interconnections that must be represented in an artificial machine. However, a subset of the brain circuitry can still be represented in an artificial machine to simulate some of the human cognitive processes.

Recently, most significant advances in the field of AI have been achieved using a perceptron, shown in Fig. 1, as the artificial model of the neuron. The output y of the neuron represents the signals sent from the axon of a biological neuron and is mathematically described by

$$y = f(W \cdot x + b).$$

The x_i inputs transmit the information to the neuron through the weights W_i which correspond to the strength of the synapses. The summation of all weighted inputs and their transformation *via* activation function f are associated with the physiological role of the neuron's cell body. The bias b represents an extra variable that remains in the system even if the other inputs are absent.

ANNs are built using perceptrons as neural primitives such that the synaptic connections are either positive or negative to mimic excitatory and inhibitory neural behaviour. A nonlinear activation function can be used to define activated and deactivated behaviours in artificial neurons. ANNs can be categorized as either feed forward (where connections between neurons do not form a cycle) or recurrent neural networks (where cycles exist).

Attempts to build fast and efficient perceptron-based ANNs have been

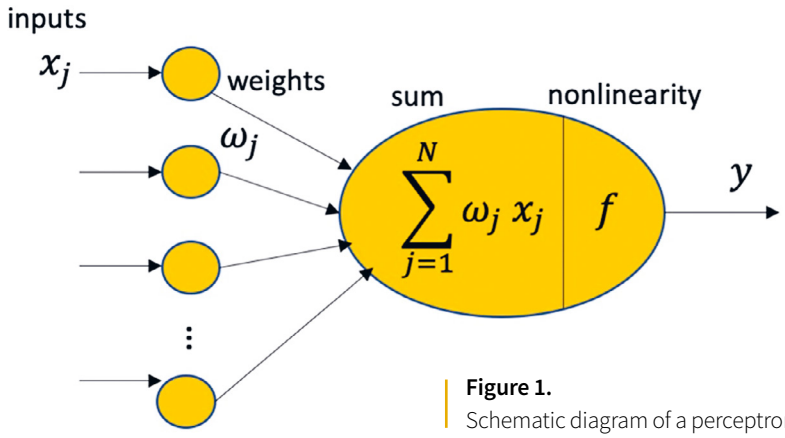
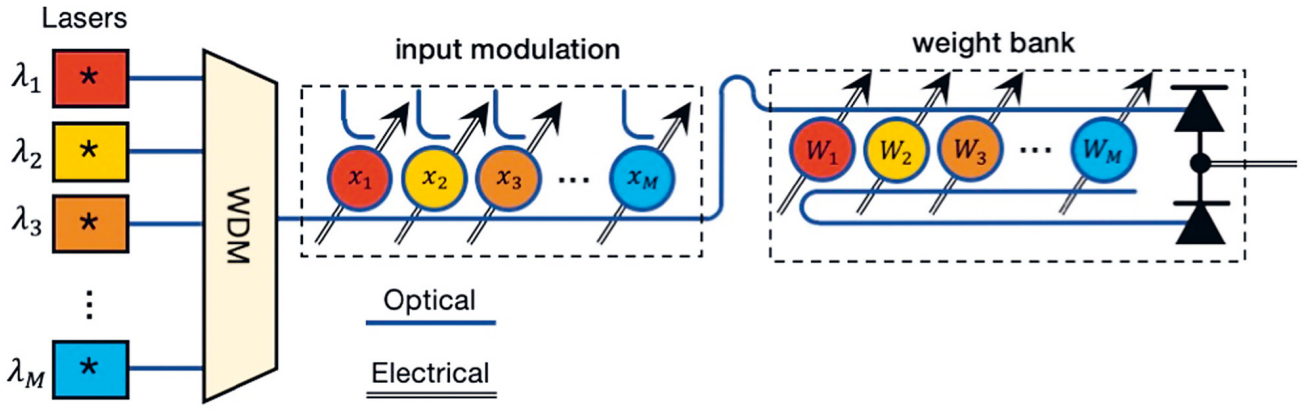


Figure 1.
Schematic diagram of a perceptron.



reported throughout recent years. An interesting computing acceleration technique consists in using hardware units to perform MAC operations at high speeds. A MAC unit performs multiplications and accumulation processes: (a+$w.x$). Multiple MAC operations can be run in parallel to perform complex operations such as convolutions and digital filters. MACs are typically used in implementations of ANNs in digital electronics [4]. Nevertheless, the serialization of the summands to perform weighted addition makes this process inefficient; consequently, chip designers are looking for alternative solutions such as full parallelism. One of the most promising technologies for this purpose is based on the photonic platform.

PHOTONIC PERCEPTRON AND MAC OPERATIONS

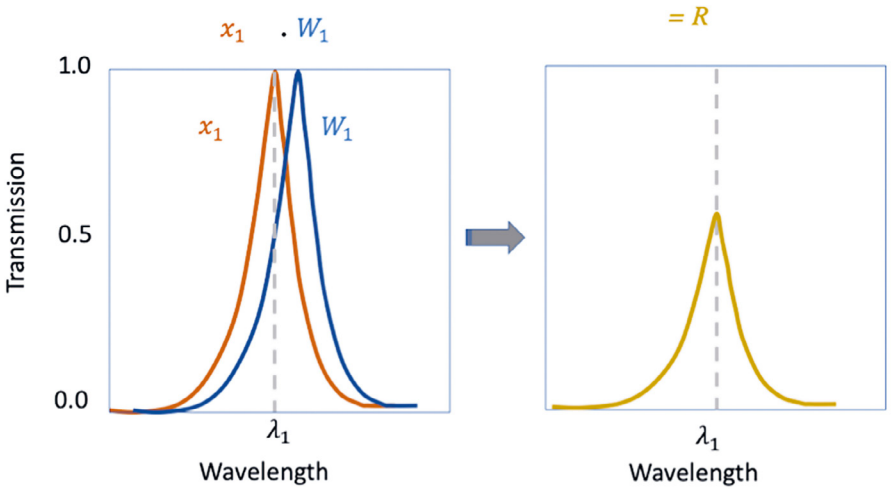
A scalable photonic architecture that implements parallel MACs can be achieved using on-chip wavelength division multiplexing (WDM) techniques [8]. This design uses microring resonators (MRRs) [9], *i.e.* photonic synapses, to encode input values and weights onto multiple wavelength signals. Tuning a given MRR on and off resonance changes the transmission of each signal through the respective filter, effectively multiplying the signal with a desired weight. An advantage of using MRRs is the ability to tune the weight values using a variety of different methods: thermally, electro-optically, or through light absorption such as phase-change or graphene materials. In this work, tuning is performed by thermally modifying the refractive index of the MRR waveguide. The application of voltage values to the heater allows us to map real-valued numbers to the device.

Figure 2.
Add-drop MRR weight bank with a balanced photodetector implementing M element-wise multipliers to perform N MAC operations in parallel.

An array of M MRRs can emulate the weighted addition of a single neuron if add-drop MRRs and a balanced photodetector are incorporated into the model, as shown in Fig. 2. In this illustration we show how to perform M MAC operations in parallel in photonics. Input values to the neuron can be mapped to voltage values V_i that tune each individual MRR(x_i). Each voltage value has a one-to-one correspondence with an MRR transmission profile T_i, and the same principle holds for weight values. The experimental implementation of this method requires the use of M lasers with different wavelengths λ_i (with i=1, … ,M) that represent M channels.

Two MRRs with different on and off resonance configurations at the same wavelength λ_1 will therefore perform element-wise multiplications, as shown in Fig. 3. Here, we show an illustration of the multiplication

Figure 3.
(a) Transmission versus wavelength curves of two different MRRs (MRR(x_1), MRR(W_1) performing element-wise optical multiplications, and (b) the product of such multiplication.



between two transmission elements x_1 and W_1, yielding the resulting value R. In Fig. 3(a), the element x_1 is tuned to have the maximum optical transmission, whereas W_1 is tuned to half the maximum. To implement x_1, MRR(x_1) is set on-resonance with λ_1 and MRR(W_1) is tuned to be half off-resonance with the same wavelength. They represent real-valued numbers 1 and 0.5, respectively. The result of such multiplication, shown in Fig. 3(b), is $R = 0.5$. A similar process is followed with the remaining sets (MRR(x_i), MRR(W_i)) for $i > 1$. Once the weighted-addition is performed using a balanced photodetector, an on-chip nonlinear function can be added by using a microring modulator.

Based on this scheme, we can design systems to solve many complex AI tasks. In the following sections, we will describe how to efficiently implement ANN training and inference on photonic chips.

Benefiting from the speed and energy advantages of photonics over traditional digital computers, the DFA training algorithm can be implemented *in situ* on silicon photonic hardware

APPLICATIONS

To implement ANNs on photonic chips, we stack N element-wise multipliers that perform weighted additions, as shown in Fig. 4. The $N \times M$ input values received from digital-to-analog converters (DACs) modulate the intensities of a group of M lasers with identical powers but unique wavelengths. These modulated inputs are sent into an array of photonic $N \times M$ weight banks (uploaded from the DACs), which then perform the multiplications for each channel. This architecture is a general representation of the multiwavelength platform as it can be used for inference, as demonstrated in [8], as well as *in situ* training.

ON-CHIP NEURAL NETWORK TRAINING

Benefiting from the speed and energy advantages of photonics over traditional digital computers, the DFA training algorithm can be implemented *in situ* on silicon photonic hardware [6]. The DFA algorithm is a supervised learning algorithm for training ANNs, where the error is propagated through fixed random feedback connections directly from the output layer to the hidden layer. The DFA algorithm has been used to train ANNs using the MNIST, Cifar-10, and Cifar-100 datasets, and yields comparable performance to the popular backpropagation training

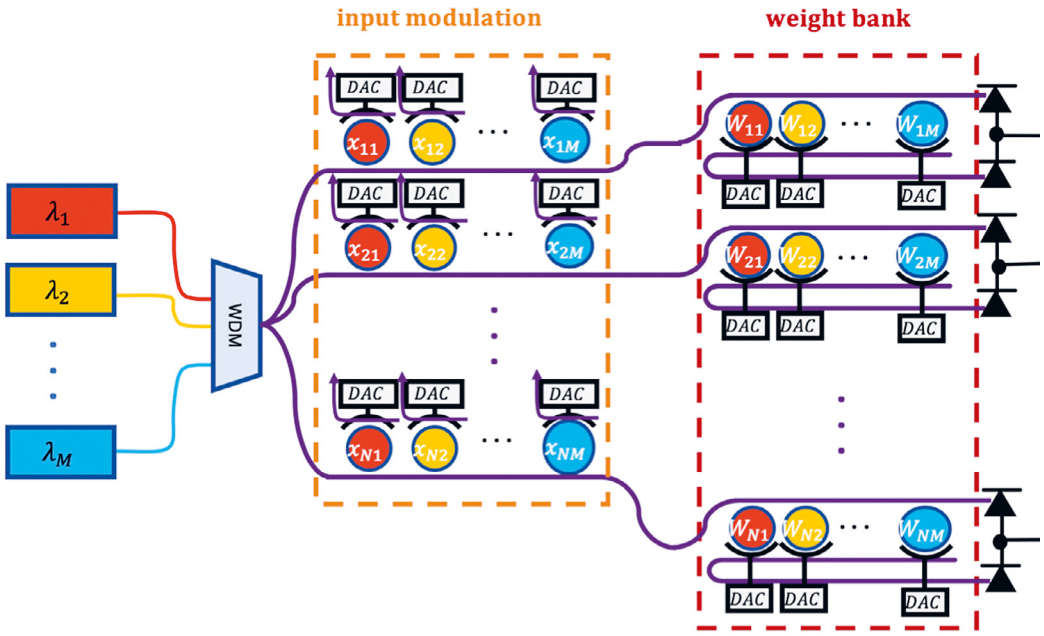


Figure 4. The input and kernel values modulate the MRRs via electrical currents proportional to those values.

algorithm [10]. A DFA photonic integrated circuit can be designed with two connected blocks with $M = 10$ and $N = 100$. This design could perform 2000 MACs per pass, enabling weight updates between two layers of 1000 neurons in 1000 passes.

LONG-SHORT TERM MEMORY NEURAL NETWORK

Similar to the DFA circuit, LSTM networks [11] can also be implemented using the multiwavelength photonic architecture [7]. An LSTM network is a recurrent architecture that offers advantages for time-series processing. Neuromorphic photonic LSTMs offer a solution to the growing demand for high-speed, high-bandwidth neural networks in time-series applications, including video processing, autonomous driving, and optical communications. The performance of the photonic LSTM for inference tasks was tested by applying the network to a simple univariate time series data problem in simulation. The simulation of this task demonstrates that even very small photonic LSTM networks performing up to 64 MACs per pass can be highly effective at performing inference tasks time series data.

CONCLUSION

Neuromorphic photonics promises exciting developments for the future of AI. In an effort to extend the bounds of digital computers for AI applications, the high bandwidth operation and full programmability of analog photonic integrated circuits can facilitate ultrafast learning and inference of ANNs. Current implementations of photonic machines face complex technical challenges that many research groups and companies have begun addressing, including the control of the processing unit and efficient memory access. Successful solutions to these problems could enable the widescale adoption of photonic processors to tackle practical AI applications.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (The MIT Press, 2016)

[2] V. Sze, Y.-H. Chen, T.-J. Yang *et al.*, Proc. IEEE **105**, 2295 (2017)

[3] P.R. Prucnal, B.J. Shastri, *Neuromorphic Photonics*, (CRC Press, 2017)

[4] M.A. Nahmias, T. Ferreira de Lima, A.N. Tait *et al.*, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2020)

[5] J. Feldmann, N. Youngblood, C.D. Wright *et al.*, Nature **569**, 208 (2019)

[6] M. J. Filipovich, Z. Guo, B. A. Marquez, *et al.*, in Proc. IEEE Photon. Conf. (IPC), paper TuA3.2 (2020)

[7] E. R. Howard, B. A. Marquez, B. J. Shastri, in Proc. IEEE Photon. Conf. (IPC), paper TuA3.2 (2020)

[8] V. Bangari *et al.*, IEEE J. Sel. Top. Quant. Electron. **26**, 7701213 (2020)

[9] A.N. Tait, T. Ferreira de Lima, E. Zhou *et al.*, Sci. Rep. **7**, 7430 (2017)

[10] A. Nøkland, Adv. Neural Inf. Process Syst. **29**, 1037 (2016)

[11] S. Hochreiter, J. Schmidhuber, Neural Comput. **9**, 173 (1997)