Integrated neuromorphic photonics

Bhavin J. Shastri, Mitchell A. Nahmias, Alexander N. Tait, Thomas Ferreira de Lima, Hsuan-Tung Peng, et al.

Bhavin J. Shastri, Mitchell A. Nahmias, Alexander N. Tait, Thomas Ferreira de Lima, Hsuan-Tung Peng, Paul R. Prucnal, "Integrated neuromorphic photonics," Proc. SPIE 10721, Active Photonic Platforms X, 107211M (19 September 2018); doi: 10.1117/12.2322182

Event: SPIE Nanoscience + Engineering, 2018, San Diego, California, United States

Invited Paper

Integrated Neuromorphic Photonics

Bhavin J. Shastri[a], Mitchell A.Nahmias[b], Alexander N. Tait[b], Thomas Ferreira de Lima[b], Hsuan-Tung Peng[b], and Paul R. Prucnal[b]

[a]Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston ON, K7L 3N6, Canada
[b]Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

ABSTRACT

Neuromorphic photonics is an emerging field at the intersection of photonics and neuromorphic engineering, with the goal of producing accelerated processors that combine the information processing capacity of neuromorphic processing architectures and the speed and bandwidth of photonics. It is motivated by the widening gap between current computing capabilities and computing needs that result from the limitations of conventional, microelectronic processors. Here, I will present these challenges, describe photonic neural-network approaches being developed by our lab and others, and offer a glimpse at this fields future.

Keywords: Multiply-accumulate (MAC), neural networks, neuromorphic computing, neuromorphic photonics, optoelectronics, photonic integrated circuits (PICs), recurrent networks, semiconductor lasers, silicon photonics, wavelength-divison multiplexing (WDM)

1. INTRODUCTION

Neuromorphic (i.e., brain-inspired) processors are widely considered as one the next frontiers in computing. The proliferation of microelectronics has enabled the emergence of next-generation industries to support emerging artificial intelligence services and high-performance computing. These data-intensive enterprises rely on continual improvements in hardware. The demand for data will continue to grow as smart gadgets multiply and become increasingly integrated into our daily lives. However, this rapidly expanding space has been subverted by a stark reality: exponential hardware scaling in digital electronics, most famously embodied in Moore's law, is fundamentally unsustainable.

Neuromorphic photonics (Fig. 1) is an emerging field at the interface of photonics and neuroscience that combines the advantages of optics and electronics to build systems with high efficiency, high interconnectivity and high information density.[1,2] Here, we will briefly look at some of the traditional challenges of photonic information processing, describe the photonic neural-network approaches being developed by our lab and others, and conclude with a future outlook of neuro-inspired photonic processing.

In the latter half of the 20th century, microprocessors faithfully adhered to Moore's law, the well-known prediction of exponentially improving performance. As Gordon Moore originally predicted in 1965, the density of transistors, clock speed, and power efficiency in microprocessors doubled approximately every 18 months for most of the past 60 years. Yet this trend began to languish over the last decade. A law known as Dennard scaling, which states that microprocessors would proportionally increase in performance while keeping their power consumption constant, has broken down since about 2006; the result has been a trade-off between speed and power efficiency. Although transistor densities have so far continued to grow exponentially, even that scaling will stagnate once device sizes reach their fundamental quantum limits in the next ten years.

One route toward resolving this impasse lies in photonic integrated circuit (PIC) platforms, which have recently undergone rapid growth. Photonic communication channels are not bound by the same physical laws as electronic ones; as a result, photonic interconnects are slowly replacing electrical wires as communication

Further author information: (Send correspondence to B.J.S.)
B.J.S.: E-mail: shastri@ieee.org

Active Photonic Platforms X, edited by Ganapathi S. Subramania, Stavroula Foteinopoulou,
Proc. of SPIE Vol. 10721, 107211M · © 2018 SPIE · CCC code: 0277-786X/18/$18
doi: 10.1117/12.2322182

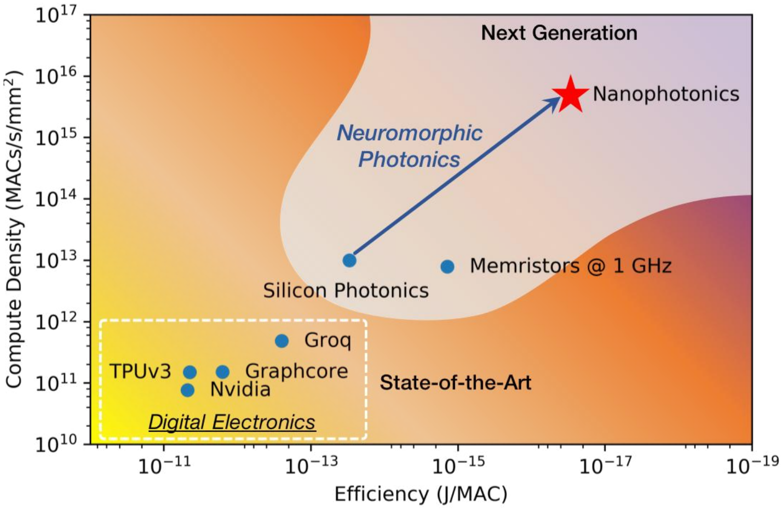

Figure 1. Comparison of specialized, deep learning digital electronic architectures with silicon photonic and nanophotonic platforms. Photonic systems can support high bandwidth densities on-chip while consuming minimal energy both transporting data and performing computations. Metrics for digital electronic architectures taken from various sources.[3–7] Silicon photonic metrics calculated assuming a modern silicon photonic platform running at 20 GHz, $N = 100$ channels with densely packed microrings. Nanophotonic metrics calculated assuming closely packed athermal microdisks.[8] ($\sim 20\mu m$ area) at 100 GHz running close to the shot noise limit.

bottlenecks worsen. PICs are becoming a key part of communication systems in data centers, where microelectronic compatibility and high-yield, low-cost manufacturing are crucial. Because of their integration, PICs can allow photonic processing at a scale impossible with discrete, bulky optical-fiber counterparts, and scalable, CMOS-compatible silicon-photonic systems are on the cusp of becoming a commercial reality. PICs have several unique traits that could enable practical, scalable photonic processing and could leap-frog the current stagnation of Moores lawlike scaling in electronic-only settings.

2. THE EMERGENCE OF PHOTONIC NEURAL NETWORKS

Instead of using digital 0's and 1's, neural networks represent information in analog signals, which can take the form of either continuous real number values, or spikes, in which information is encoded in the timing between short pulses. Rather than abiding by a sequential set of instructions, neurons process data in parallel and are programmed by the connections between them (Fig. 2). The input into a particular neuron is a linear combination—also referred to as weighted sum—of the output of other neurons. These connections can be weighted with negative and positive values, respectively, which are called "inhibitory" and "excitatory" synapses. The weight is therefore represented as a real number, and the interconnection network can be expressed as a matrix.

Photonics is a promising technology to implement neural networks (Fig. 2). The greatest computational burden in neural networks lies with the interconnectivity: in a system with N neurons, if every neuron can communicate with every other (plus itself), this results in N^2 connections. Just one more neuron adds N more connections, which can be prohibitive if N is large. Photonic systems could address this problem in two ways: 1) waveguides can boost interconnectivity by carrying many signals at the same time through optical multiplexing, and 2) low-energy, photonic operations can reduce the computational burden of performing linear functions such as weighted sum. For example, by associating each node with a color of light, a network could support N additional connections without necessarily adding any physical wires.

We can understand this better through the example of a multiply-accumulate (MAC) operation. Each such operation represents a single multiplication, followed by an addition. Since, mathematically, MAC operations comprise dot products, matrix multiplications, convolutions and Fourier transforms, they underlie much of high-performance computing. They also constitute the most costly operations in both hardware-based neural networks

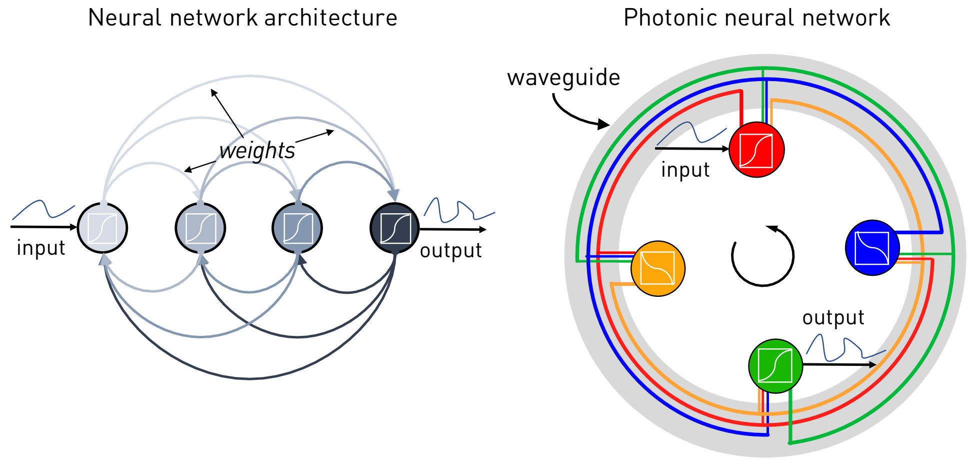

Figure 2. Photonic neural nets (right) can solve the interconnect bottleneck by using one waveguide to carry signals from many connections (easily $N^2 \sim 10,000$) simultaneously.

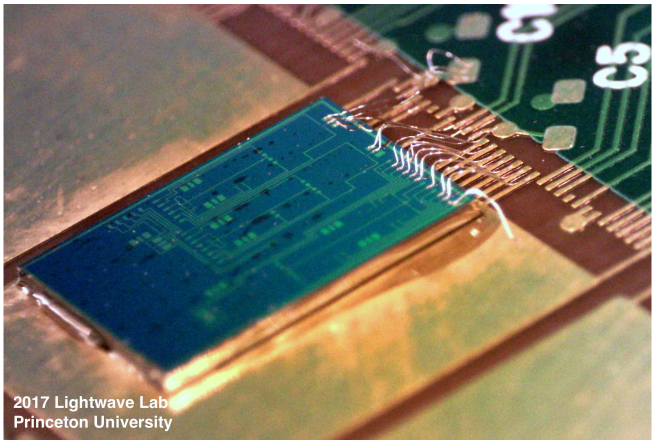

Figure 3. A laser neural network being tested at Princeton University.

and machine-learning algorithms. In the digital domain, MACs occur in a serial fashion, which means that the time and energy costs increase with the number of inputs.

In contrast, passive lightwave devices, such as wavelength-sensitive filters, do not inherently dissipate energy and can efficiently perform such operations in parallel. They can therefore greatly enhance high performance computing, especially systems that rely on matrix multiplication. A comparison of the potential speed and efficiency of photonic based systems is shown in Fig 1. In addition, reprogrammability is possible with tunable photonic elements. These advantages have motivated researchers to investigate a number of photonic neural models that exhibit a large range of interesting properties.

3. PHOTONIC NEURON IMPLEMENTATIONS

Researchers have engineered dynamical lasers to resemble the biological behavior of neurons;[9] an example of an integrated system currently under investigation at Princeton is shown in Fig. 3. Laser neurons are capable of operating approximately 100 million times the speed of their biological counterparts, owing to the speed of optoelectronic physics over biochemical interactions. They represent neural spikes via optical pulses by operating under a dynamical regime called "excitability." Excitability is a behavior in feedback systems in which small inputs that exceed some threshold cause a major excursion from equilibrium, which in the case of a laser neuron, releases an optical pulse. This event is followed by a recovery back to equilibrium, or refractory period.

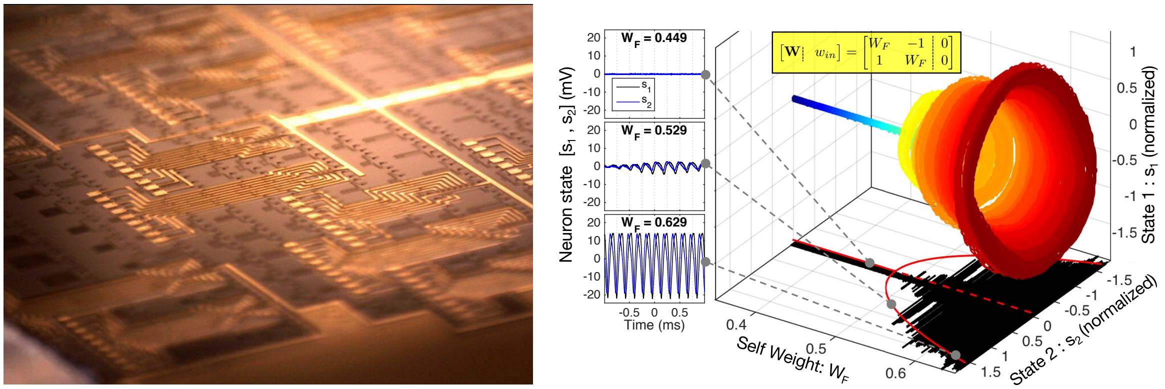

Figure 4. Left: picture of a silicon photonic broadcast-and-weight neural network platform. Right: Time traces of a modulator-class, tunable 2× recursive neuron network. Increasing the weight parameter W_F beyond a critical value results in a Hopf bifurcation and the formation of a limit cycle. This shows that the network has controllable, stable dynamics. Black shadow: average experimental amplitudes; solid red curve: corresponding fit model; dotted red line: unstable branch. Reproduced from Tail *et al.*, *Sci. Rep.* **7**, 7430 (2017).[12] Licensed under Creative Commons Attribution License. (CC BY).

We discovered[10] a theoretical link between the dynamics of semiconductor lasers and a common neuron model used in computational neuroscience, and demonstrated how a laser with an embedded graphene section could effectively emulate such behavior.[11] Building from these results, a number of researchers have fabricated, tested, and proposed a variety of laser neurons with various feedback conditions.[9] These include two-section models in semiconductor lasers, photonic crystal nanocavities, polarization sensitive vertical cavity lasers, lasers with optical feedback or optical injection, and linked photodetector-laser systems with receiverless connections or resonant tunneling.

A recently demonstrated[12] approach based on optical modulators has been investigated recently that has the potential to exhibit much lower conversion costs from one processing stage to another. In addition, it would be fully integrated systems on silicon photonic platforms.

4. SCALABLE PHOTONIC NEURAL NETWORKS

Recently, researchers have investigated interconnection protocols that can tune to any desired network configuration. Arbitrary weights allow a wide array of potential applications based on classical neural networks. There are several notable approaches in the literature that use complementary physical effects in this regard.

Broadcast-and-weight: A neural network architecture called “broadcast and weight” uses groups of tunable filters to implement weights on signals encoded onto multiple wavelengths.[13] Tuning a given filter on and off resonance changes the transmission of each signal through that filter, effectively multiplying the signal with a desired weight. The resulting weighted signals travel into a photodetector, which can receive many wavelengths in parallel to perform a summing operation. Broadcast and weight takes advantage of the enormous information density available to on-chip photonics through the use of optical multiplexing, and is compatible with a number of laser neuron models. Filter-based weight banks have also been investigated both theoretically and experimentally in the form of closely packed microring resonator (MRR) filters, prototyped in a silicon photonic platform. A fully integrated superconducting optoelectronic network was recently proposed to offer unmatched energy efficiency.[14] While based on an exotic superconducting platform, the interconnect architecture could be compatible with broadcast-and-weight.

Broadcast-and-weight networks offer a compact way to instantiate large networks of neurons (Fig. 4). As an illustrative example, conservatively assuming each MRR occupies 250 μm^2 in area, an implementation of SqueezeNet (421,098 parameters)[15] would take up approximately ~1 cm^2. Thus, as one looks towards systems that rival the size of current software neural networks, scaling beyond the channel count limit becomes critical. This can be achieved by chaining multiple broadcast waveguides together via interfacing PNNs. In multi-broadcast systems, the number of wavelength channels only limits the fan-in per processor rather than the total

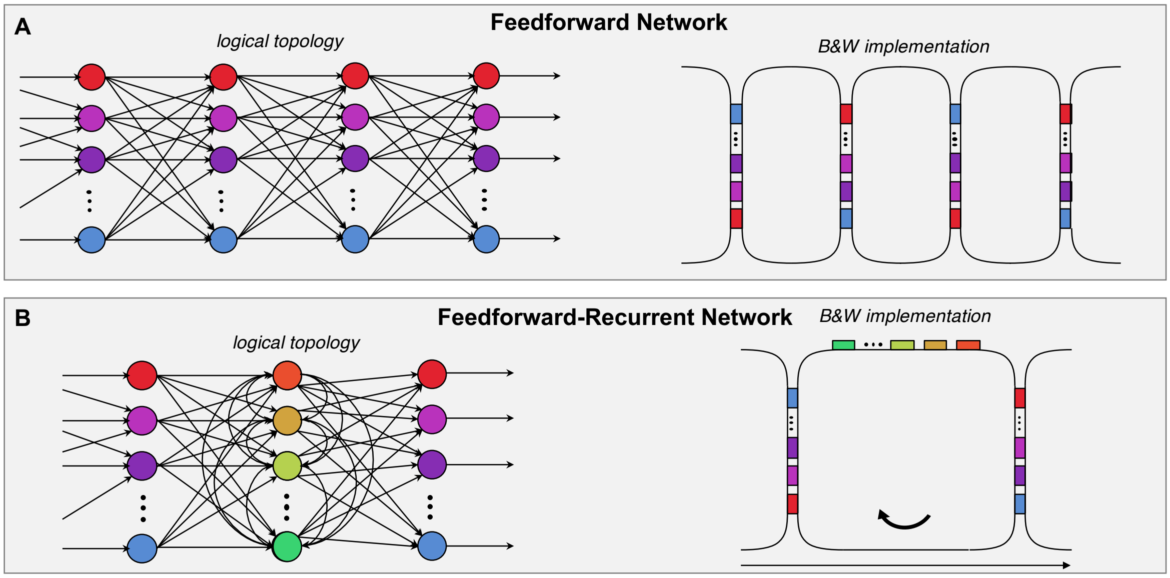

Figure 5. The broadcast-and-weight protocol can scale beyond the all-to-all N wavelength limit through the use of interfacial nodes. Examples of how A) feedforward networks can be constructed using interfacial PNNs that connect between broadcast mediums, and B) chains of recurrent networks can be constructed using interfacing nodes between self-connected broadcast mediums.

size of the network. However, one must impose substructure on the network since there are limitations on the number of photonic neurons that can interface between waveguides. Several examples of useful topologies are shown in Fig. 5, wherein each topology may be better suited for different applications. Beyond these base configurations, it is possible to create more complex, small-world-like hierarchical networks as described in.[13]

Coherent: A "coherent" approach utilizes destructive or constructive interference effects in optical interferometers to implement a matrix-vector operation of incoming signals.[16] There is no need to convert from the optical domain to the electrical domain; interfacing such systems with photonic, nonlinear nodes (i.e., based on the Kerr effect) could allow for energy efficient, passive all-optical processors. However, the coherent approach is limited to only one wavelength and requires devices that are much larger than tunable filters, limiting the information density of the approach in its current form. In addition, all-optical interconnects must grapple with both amplitude and phase and there is still no proposed solution to prevent phase noise accumulation from one stage to another. Nonetheless, the investigation of large-scale networking schemes is a promising direction for the integration of various technologies in the field towards highly scalable on-chip photonic systems.

Reservoir computing: A contrasting approach to tunable neural networks, "reservoir computing," extracts useful information from a fixed, possibly nonlinear system of interacting nodes.[17] Reservoirs require far fewer tunable elements than neural network models to run effectively, making them less challenging to implement in hardware; however, they cannot be easily programmed. These systems have utilized optical multiplexing strategies in both time and wavelength. Experimentally demonstrated photonic reservoirs have displayed state-of-the-art performance in benchmark classification problems, such as speech recognition.

5. DISCUSSION

Although it remains to be seen in what ways photonic processing systems will complement microelectronic hardware, current technological developments point in a promising direction. For example, the fixed cost of electronic to photonic conversion is no longer as energetically unfavorable: a modern silicon photonic link can transmit a photonic signal using only femtojoules of energy per bit of information, while thousands of femtojoules of energy are consumed per operation in even the most efficient digital electronic processors, including IBM's TrueNorth cognitive computing chip and Google's tensor processing unit. This figure will improve as optoelectronic devices are scaled in performance. New modulators or lasers based on plasmonic localization, graphene modulation or nanophotonic cavities have the potential to increase this efficiency. The next generation of photonic devices could potentially consume only hundreds of attojoules of energy per time slot, allowing analog photonic processors to consume even less per operation.

There are many applications of photonic neural network technologies, especially in light of the developments mentioned above. For one, photonic systems can act as a co-processor to perform linear operations—including multiply-accumulate operations, fourier transforms, and convolutions—by implementing them in the photonic domain, potentially decreasing the energy consumption and increasing the throughput of signal processing, high performance computing and artificial intelligence algorithms. This could be a major boon for datacenters, which are increasingly dependent on such operations and have consistently doubled their energy consumption every four years.

Secondly, photonic processors have unmatched speeds and latencies, which make them well-suited for specialized applications requiring either real-time response times or fast signals. One example is a front-end processor in radio frequency transceivers. As the wireless spectrum becomes increasingly overcrowded, the use of large, adaptive phased-array antennas that receive many more radio waves simultaneously may soon become the norm. Photonic neural networks could perform complex statistical operations to extract important data, including the separation of mixed signals or the classification of recognizable radio frequency signatures. A second example is in low-latency, ultrafast control systems. It is well known that recurrent neural networks can solve various problems that involve minimizing or maximizing some known function. A process method known as "Hopfield optimization" requires the solution to such a problem during each step of the algorithm, and could utilize the short convergence times of photonic networks for nonlinear optimization.

Just as fiber optics once rendered copper cables obsolete for long-distance communications, neuromorphic photonic processing has the potential to one day usher a paradigm shift in computing to create a smarter, more efficient world.

REFERENCES

[1] Prucnal, P. R. and Shastri, B. J., [*Neuromorphic Photonics*], CRC Press, Taylor & Francis Group, Boca Raton, FL, USA (2017).

[2] Peng, H. T., Nahmias, M. A., de Lima, T. F., Tait, A. N., and Shastri, B. J., "Neuromorphic photonic integrated circuits," *IEEE Journal of Selected Topics in Quantum Electronics* **24**, 1–15 (Nov 2018).

[3] "Groq," (November 2017).

[4] Teich, P., "Tearing apart google's TPU 3.0 AI coprocessor," (May 2018).

[5] Smith, R., "Nvidia Volta unveiled: GV100 GPU and Tesla V100 accelerator announced," (May 2017).

[6] Knowles, S., "Scalable silicon compute," (December 2017).

[7] Wijesinghe, P., Ankit, A., Sengupta, A., and Roy, K., "An all-memristor deep spiking neural network: A step towards realizing the low power, stochastic brain," *arXiv preprint arXiv:1712.01472* (2017).

[8] Timurdogan, E., Sorace-Agaskar, C. M., Sun, J., Hosseini, E. S., Biberman, A., and Watts, M. R., "An ultralow power athermal silicon modulator," *Nature Communications* **5**, 4008 (2014).

[9] Prucnal, P. R., Shastri, B. J., de Lima, T. F., Nahmias, M. A., and Tait, A. N., "Recent progress in semiconductor excitable lasers for photonic spike processing," *Advances in Optics and Photonics* **8**, 228–299 (Jun 2016).

[10] Nahmias, M. A., Shastri, B. J., Tait, A. N., and Prucnal, P. R., "A Leaky Integrate-and-Fire Laser Neuron for Ultrafast Cognitive Computing," *IEEE Journal of Selected Topics in Quantum Electronics* **19**(5) (2013).

[11] Shastri, B. J., Nahmias, M. A., Tait, A. N., Rodriguez, A. W., Wu, B., and Prucnal, P. R., "Spike processing with a graphene excitable laser," *Scientific Reports* **6**, 19126 EP – (01 2016).

[12] Tait, A. N., de Lima, T. F., Zhou, E., Wu, A. X., Nahmias, M. A., Shastri, B. J., and Prucnal, P. R., "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific Reports* **7**(1), 7430 (2017).

[13] Tait, A. N., Nahmias, M. A., Shastri, B. J., and Prucnal, P. R., "Broadcast and weight: An integrated network for scalable photonic spike processing," *Journal of Lightwave Technology* **32**, 3427–3439 (Nov 2014).

[14] Shainline, J. M., Buckley, S. M., Mirin, R. P., and Nam, S. W., "Superconducting optoelectronic circuits for neuromorphic computing," *Physical Review Applied* **7**, 034013 (Mar 2017).

[15] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K., "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360* (2016).

[16] Shen, Y., Harris, N. C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., and Soljačić, M., "Deep learning with coherent nanophotonic circuits," *Nature Photonics* **11**, 441 EP – (06 2017).

[17] der Sande Guy, V., Daniel, B., and C., S. M., "Advances in photonic reservoir computing," **6**, 561 (July 2017).