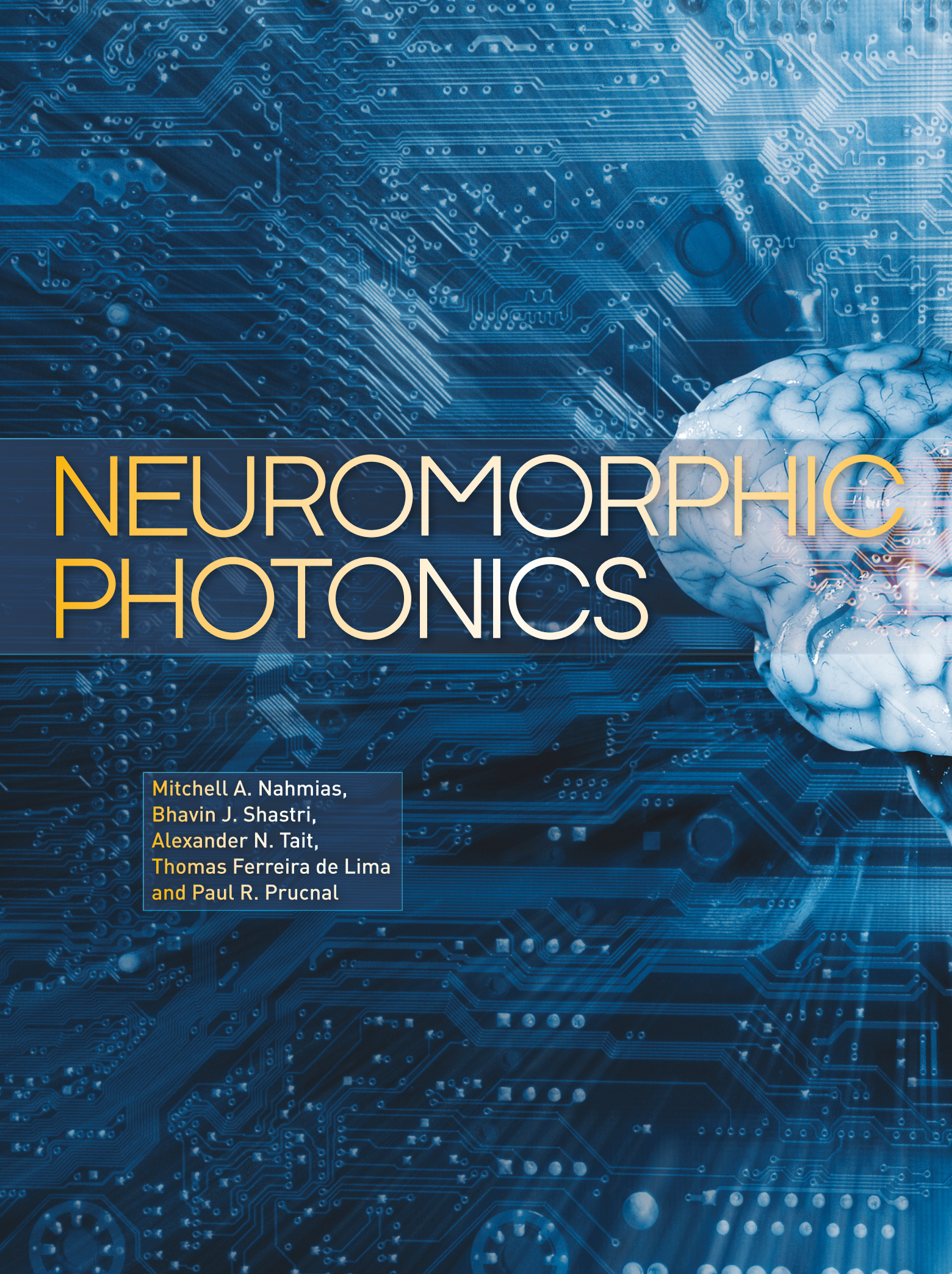


NEUROMORPHIC PHOTONICS

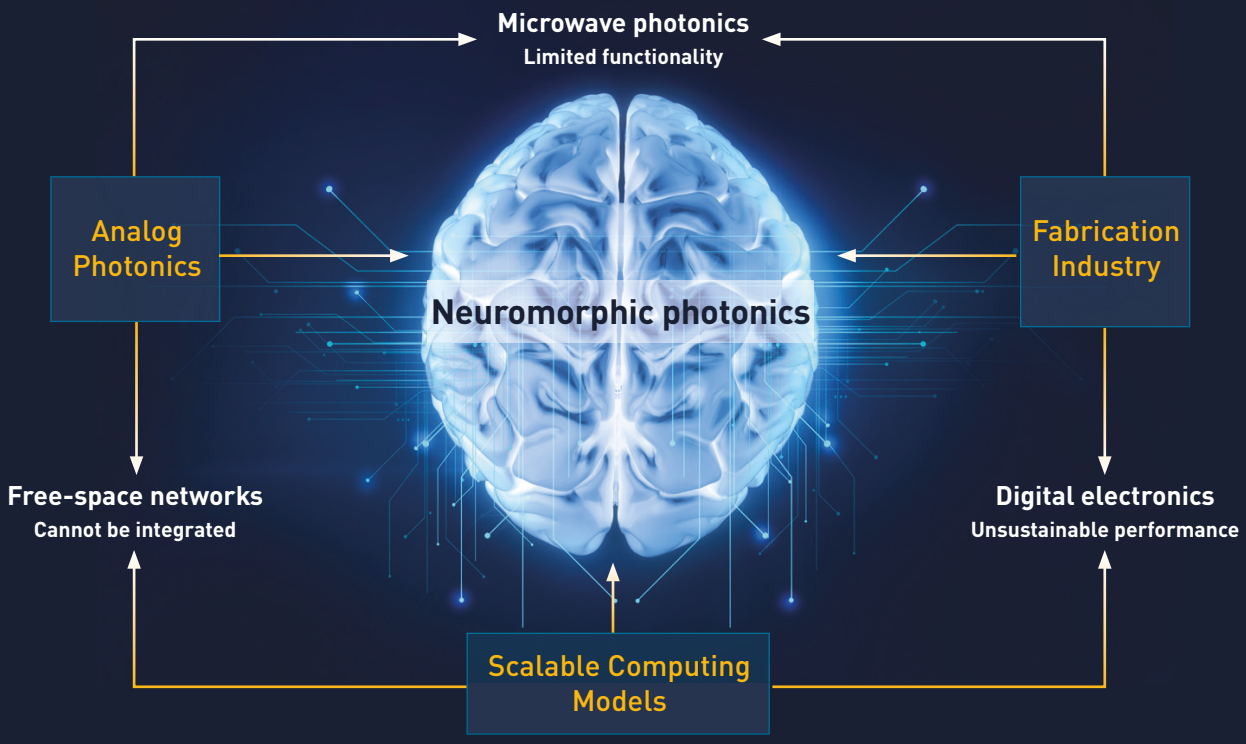
Mitchell A. Nahmias,
Bhavin J. Shastri,
Alexander N. Tait,
Thomas Ferreira de Lima
and Paul R. Prucnal

Getty Images

Photonic neural networks have the potential to revolutionize the speed, energy efficiency and throughput of modern computing—and to give Moore's law–style scaling a new lease on life.

SYNERGISTIC APPROACH

Neuromorphic photonics uses modern fabrication techniques to implement efficient, scalable analog photonics operations.



In an age overrun with information, the ability to process vast volumes of data has become crucial. 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—and the demand for data will continue to grow as smart gadgets multiply and become ever more integrated into our daily lives. Unfortunately, however, those prospects are running up against a stark reality: the exponential hardware scaling in digital electronics, most famously embodied in Moore's law, is fundamentally unsustainable.

This situation suggests that the time is ripe for a radically new approach: neuromorphic photonics. An emerging field at the nexus of photonics and neuroscience, neuromorphic photonics combines the advantages of optics and electronics to build systems with high efficiency, high interconnectivity and high information density. In the pages that follow, we take a 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 offer a glimpse at the future outlook for this emerging field.

Moving beyond Moore

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 bottlenecks worsen. PICs are becoming a key part of communication systems in data centers, where

1047-6938/18/01/34/8-$15.00 © OSA

Neuromorphic photonics combines the advantages of optics and electronics to build systems with high efficiency, high interconnectivity and high information density.

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 leapfrog the current stagnation of Moore's law–like scaling in electronic-only settings:

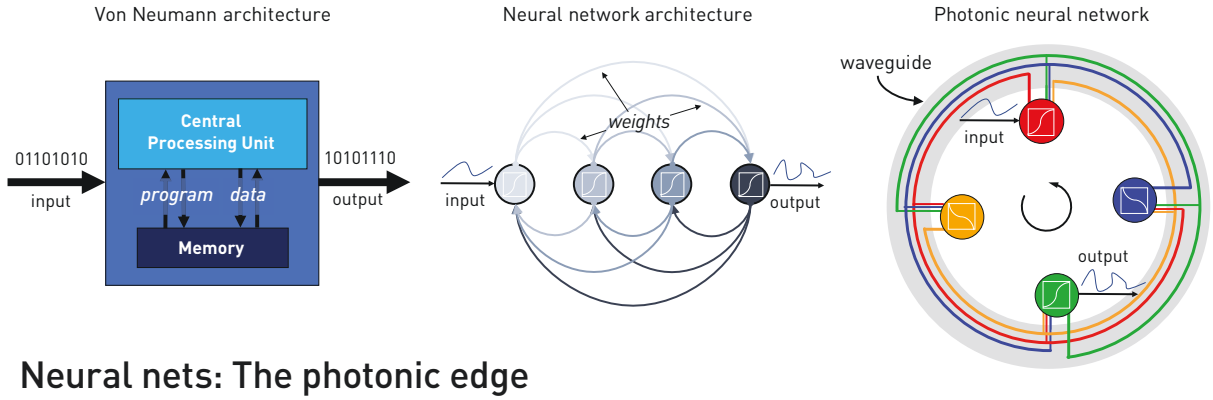
Speed. Electronic microprocessor clock rates cannot exceed about four GHz before hitting thermal-dissipation limits, and parallel architectures, such as graphic processing units, are limited to even slower timescales. In contrast, each channel in a photonic system, by default, can operate at upwards of twenty gigahertz to support fiber optic communication rates.

Information density. Paradoxically, despite the large sizes of on-chip photonic devices—whose lower bound on size must exceed the wavelength of the light that travels through them—PICs can pack orders of magnitude more information in every square centimeter. One reason is that photonic signals operate much faster, thereby shuffling much more data through the system per second. Another is that lightwaves exhibit the superposition property, which allows for optical multiplexing: waveguides can carry many signals along different wavelengths or time slots simultaneously without taking up additional space. This combination enables an enormous amount of information—easily more than one terabyte per second—to flow through a waveguide only half a micron wide.

Energy efficiency. Photonic operations have the potential to consume orders of magnitude less power than digital approaches. This property comes from so-called linear photonic operations (that is, those that can be described using linear algebra). Transmission elements are sometimes considered to dissipate no energy; however, it always takes energy to generate, modulate and receive light signals. Nonetheless, the lack of a fundamental energy cost per operation means that photonic processors may not be subject to the unfavorable scaling laws that have stymied further performance returns in electronic systems.

Photonic signal processing

Optical signal processing has a rich history, but optical systems have had difficulty achieving scalability in computing. Extensive research has focused on implementing optical-computing operations using both digital bits and continuous-valued analog signals. Concepts for neuro-inspired photonic computing originally



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^2~10,000) simultaneously.

envisioned systems that used vertically oriented light sources or spatial light modulators together with free-space holographic routing. Many researchers imagined that an optical computer would consist of a 3-D holographic cube programmed to route signals between arrays of LEDs.

Although optical logic devices later developed into the switches and routers that form today's telecommunications infrastructure, optical computing did not achieve the same level of success. Researchers realized that the scaling laws for electronic components could continue to address the bottlenecks in traditional processors for many years to come. The ceaseless march of Moore's law meant that, while optical computing systems might outperform electronics in the short term, microprocessors would eclipse them in several years.

A close look at the hardware reveals that the past challenges of optical computing—and, particularly, optical neural computing—lay chiefly in a few factors: the continued favorable scaling of electronic devices, the packaging difficulties associated with free-space coupling and holographic interconnects, and the difficulty in shrinking optical devices. Now, about 30 years later, the landscape has changed tremendously. With Moore's law confronting fundamental limitations, the scaling of electronics can no longer be taken for granted. Meanwhile, large-scale integration techniques are starting to emerge in photonics, driven by telecommunication applications and a market need for increased information flow both between and within processors.

These changes have led to an explosion in PICs, which are already finding their way into fast Ethernet switches in servers and data centers. Microwave photonics are also emerging as a contender for radio-frequency applications, now enabled by the low cost of microchip photonic integrated components. Researchers have implemented digital photonic devices in various technologies, including fibers, waveguides, semiconductor devices and resonators.

Both the analog and the digital approaches to optical computing, however, still face challenges. Increasing the number of analog operations leads to noise and degrades signal integrity, limiting the potential complexity of optical processors. And, while digital systems filter out noise during every step and can fix errors after they occur—making it easy for engineers to design complex systems with many interacting components—the high scaling cost of digital photonic devices makes this approach both prohibitively expensive and impractical.

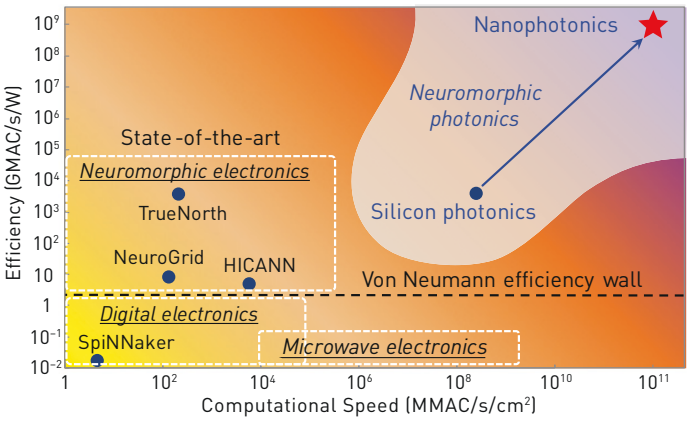
Photonic neural networks

Neural network approaches represent a hybrid between the purely digital and analog approaches, allowing for more efficient processors that are both less resource-intensive and robust to noise. But what *is* a neural network?

Most modern microprocessors follow the so-called von Neumann architecture, in which machine instructions and data are stored in memory and share a central communication channel, or bus, to a processing unit. Instructions define a procedure to operate on data, which is continually shuffled back and forth between memory and the processor.

Neural networks function quite differently. Individually, neurons can perform simple operations such as adding inputs together or filtering out weaker signals. In groups, however, they can implement far more complex operations through the formation of 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 of 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.

The input into a particular neuron is a linear combination—also referred to as a weighted addition—of



Electronic vs. photonic neural nets

Neuromorphic 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, the University of Heidelburg's HICANN), as well as advanced digital electronic systems (such as the University of Manchester's SpiNNaker).

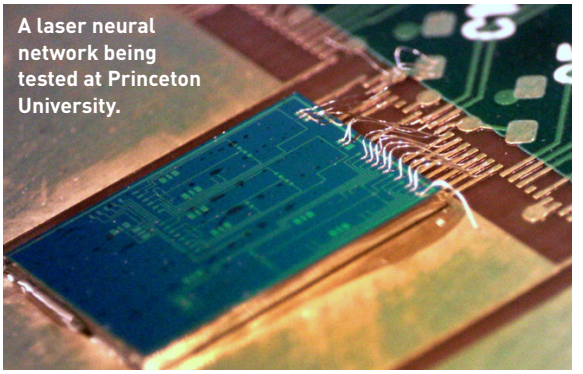
Rather than abiding by a sequential set of instructions, neurons process data in parallel and are programmed by the connections between them.

the output of other neurons. These connections can be weighted with negative and positive values, respectively, which are called (borrowing the language of neuroscience) inhibitory and excitatory synapses. The weighting is therefore represented as a real number, and the interconnection network can be expressed as a matrix.

Photonics appears to be an ideal technology with which to implement neural networks. The greatest computational burden in neural networks lies in the interconnectivity: in a system with N neurons, if every neuron can communicate with every other neuron (plus itself), there will be N^2 connections. Just one more neuron adds N more connections—a prohibitive situation if N is large. Photonic systems can address this problem in two ways: waveguides can boost interconnectivity by carrying many signals at the same time through optical multiplexing; and low-energy, photonic operations can reduce the computational burden of performing linear functions such as weighted addition. 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 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. In addition, reprogrammability is possible with tunable photonic elements. These advantages have motivated researchers to investigate a variety of photonic neural models that exhibit a range of interesting properties.



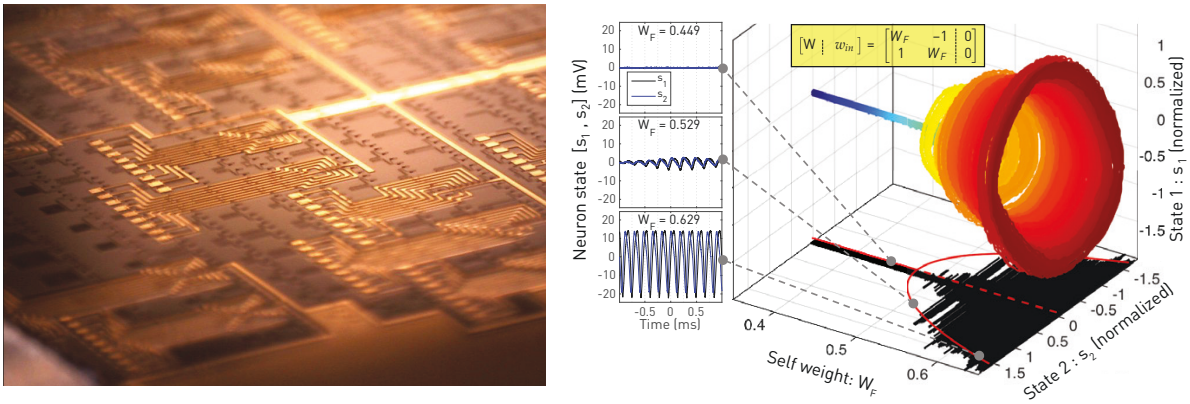
A laser neural network being tested at Princeton University.

Princeton University Lightwave Lab, 2017

A spectrum of implementations

One such photonic neural model, currently under investigation in our lab, involves engineering dynamical lasers to resemble the biological behavior of neurons. Laser neurons, operating optoelectronically, can operate at approximately 100 million times the speed of their biological counterparts, which are rate-limited by biochemical interactions. These lasers 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, the so-called refractory period.

We have found a theoretical link between the dynamics of semiconductor lasers and a common neuron model used in computational neuroscience, and have demonstrated how a laser with an embedded graphene section could effectively emulate such behavior. Building from these results, a number of research groups have fabricated, tested and proposed laser neurons with various feedback conditions. 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



Left: A photonic neural network that can be implemented in silicon photonics. Right: The on-chip system with modulator neurons displays a characteristic oscillation called a Hopf bifurcation, which confirms the presence of an integrated neural network. Princeton University Lightwave Lab, 2017/ A. Tait et al., Sci. Rep. **7**, 7430 (2017).

resonant tunneling. A recently demonstrated approach based on optical modulators has the potential to exhibit much lower conversion costs from one processing stage to another, and to be fully integrated on silicon-photonic platforms.

Toward scalable networks

Researchers have lately 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. Several notable approaches use complementary physical effects in this regard.

Broadcast-and-weight. A broadcast-and-weight neural network architecture, demonstrated by our group at the Princeton Lightwave Lab, uses groups of tunable filters to implement weights on signals encoded onto multiple wavelengths. 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 filters, prototyped in a silicon-photonic platform. And the interconnect architecture of a fully integrated superconducting optoelectronic network recently proposed by scientists at the U.S. National Institute of Standards and Technology—and said to offer potentially unmatched energy efficiency—could be compatible with broadcast-and-weight.

Coherent. A coherent approach, which uses destructive or constructive interference effects in optical interferometers to implement a matrix-vector operation on incoming signals, was recently demonstrated by a research team led by Marin Soljačić and Dirk Englund and at the Massachusetts Institute of Technology, USA. In such an architecture there is no need to convert from the optical domain to the electrical domain; hence, interfacing a coherent system with photonic, nonlinear nodes (for example, based on the Kerr effect) could in principle allow for energy efficient, passive all-optical processors.

The coherent approach is, however, limited to only one wavelength, and requires devices much larger than tunable filters, which puts a cap on the information density that the approach can achieve in its current form. In addition, all-optical interconnects must grapple with both amplitude and phase, and no solution has yet been proposed 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 being pursued by a number of labs, reservoir computing extracts useful information from a fixed, possibly nonlinear system of interacting nodes. Reservoirs require far fewer tunable elements than neural-network models to run effectively, making

Neuromorphic photonic processing has the potential to one day usher in a paradigm shift in computing—creating a smarter, more efficient world.

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.

Marching ahead

It remains to be seen in what ways photonic processing systems will complement microelectronic hardware, but current technological developments look promising. For example, the fixed cost of electronic-to-photonic conversion is no longer as energetically unfavorable as in the past. A modern silicon-photonic link can transmit a photonic signal using only femtojoules of energy per bit of information, whereas 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.

The comparisons should get better still as performance scaling in optoelectronic devices continues to improve. New modulators or lasers based on plasmonic localization, graphene modulation or nanophotonic cavities have the potential to increase efficiency. The next generation of photonic devices could potentially consume only hundreds of attojoules of energy per time slot, allowing analog photonic MAC-based processors to consume even less per operation.

In light of these developments, photonic neural networks could find a place in many applications. These systems can act as a coprocessor for performing computationally intense linear operations—including MACs, 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 boon for data centers, which increasingly depend on such operations and have consistently doubled their energy consumption every four years.

Photonic processors also 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 radiofrequency signatures.

Still another application example lies in low-latency, ultrafast control systems. It's well understood that recurrent neural networks can solve various problems that involve minimizing or maximizing some known function. A processing 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.

Fiber optics once rendered copper cables obsolete for long-distance communications. Neuromorphic photonic processing has the potential to one day usher in a similar paradigm shift in computing—creating a smarter, more efficient world. OPN

Mitchell A. Nahmias, Bhavin J. Shastri, Alexander N. Tait, Thomas Ferreira de Lima and Paul R. Prucnal (prucnal@princeton.edu) are with the Department of Electrical Engineering, Princeton University, Princeton, N.J., USA.

References and Resources

- P. Prucnal and B. Shastri. *Neuromorphic Photonics* (CRC Press, 2017).
- M. Nahmias et al. J. Sel. Top. Quantum Electron. **19**, 1800212 (2013).
- A. Tait et al. J. Lightwave Technol. **32**, 4029 (2014).
- B. Shastri et al. Sci. Rep. **5**, 19126 (2015).
- P. Prucnal et al. Adv. Opt. Photon. **8**, 228 (2016).
- A. Tait et al. Sci. Rep. **7**, 7430 (2017).
- Y. Shen et al. Nat. Photon. **11**, 441 (2017).
- J.M. Shainline et al. Phys. Rev. Appl. **7**, 034013 (2017).
- G. Van der Sande et al. Nanophotonics, **6**, 561 (2017).