Emergence of Neuromorphic Photonics

B. J. Shastri, T. Ferreira de Lima, A. N. Tait, M. A. Nahmias, and P. R. Prucnal

Department of Electrical Engineering, Princeton University, Princeton NJ, 08544 USA

shastri@ieee.org, prucnal@princeton.edu

Abstract: Digital electronics are becoming the limiting factor in signal processing areas such as radio. Photonic integration offers new potential for moderate-scale photonic systems. Neuromorphic Photonics bridges integrated photonic physics and neural models of information processing.

© 2017 Optical Society of America

OCIS codes: (070.1170) Analog optical signal processing; (200.4700) Optical neural systems; (250.5300) Photonic integrated circuits.

The success of digital electronics has created a data-hungry consumer society, which in turn, reinvested in more capable, faster and cheaper machines. For decades, the transistor count of CPUs doubled every two years, a trend which became known as Moore's law. Microprocessor clock rates also increased exponentially, but current leakage in nanometric nodes became prevalent, causing a halt to this growth at about 4 GHz [1]. At the same time, the past decade has seen the breakdown of Dennard scaling [2]—the power density of microelectronic chips no longer stays constant as they get denser, i.e. smaller transistors do not consume less power. The recent shift to multi-core scaling alleviated these constraints, but the breakdown of Dennard scaling has limited the number of cores than can simultaneously be powered on with a fixed power budget and heat extraction rate—giving rise to the *dark silicon* phenomenon [3]. Projections for the 8 nm node indicate that over 50% of the chip will be *dark* [3]. Fundamentally, these issues can be traced to two primary physical bottlenecks: the bandwidth limitations of metal interconnects, and the energy consumption–and subsequently, heat generation of digital switching [4]. In summary, operating speed and power efficiency of CPUs have reached physical barriers that cannot be addressed through Dennard scaling. Consequently, this has opened up new opportunities in unconventional information processing architectures, which include an array of different processing modalities [5].

Respecting power budgets is now a top priority for digital processors. Data centers, Wi-Fi routers and Internet traffic represent a tremendous electric energy consumption. Current trends indicate a shift of electricity usage from consumer device use to network and data centers [6, 7]. In the worst case scenario, at the rate at which societal consumption and production of data is growing, it is predicted that fixed-access networks (Wi-Fi and LAN) and data centers will consume up to 33% of world's energy use [7].

To counter that trend, power-aware large-scale integration techniques in photonics are just emerging, being pushed forward by data communication applications and a market need for increased information flow between processors, both on the macro and micro scale [6, 8]. This has led to an explosion in photonic integrated circuit (PICs), which are already finding their way into fast ethernet switches for servers and supercomputers, and will likely emerge in more traditional processor architectures as electronic interconnects fail to keep up with data volume. The average energy efficiency of the world's fastest supercomputers lies in the order of 1 nJ=FLOP [9], where FLOP stands for floating-point operation, a standard computing unit. In green data centers and high-performance computers, there is an urgent need for unconventional, special-purpose co-processors with efficiencies beyond 1 nJ=FLOP, with a caveat: these co-processors must operate at the same throughput handled by the high-speed digital and analog circuits it interfaces with, so they do not become a bottleneck.

This efficiency level is not fundamentally impossible. In fact, the human brain is estimated to being able to compute an amazing 10^{20} MAC=s using only 20 W of power [10] (MAC: multiply and accumulate operation, similar to FLOP but more appropriate for digital signal processors). It does this with 10^{11} neurons with spike firing rate of 1 Hz but with a large number of interconnects per neuron (10^4), highlighting the importance of distributed processing. The calculated computational efficiency for the brain is therefore 9 orders of magnitude beyond that of current supercomputers ($<$ aJ=MAC). *Neuromorphic computing* offers hope to building large-scale "bio-inspired" hardware for specialized processing while attempting computational efficiencies toward those of a human brain.

We will review our recent progress in neuromorphic photonics research [11–16], focusing especially on integrated photonic devices. An elegant parallel between neural networks and optoelectronic devices such as excitable lasers

can be established and exploited for processing. We will introduce the concept of a *photonic neuron*, followed by a discussion on its feasibility. Finally, we will also present a scalable neurmorphic networking architecture that efficiently channelizes the spectrum of an integrated waveguide.

Photonics has revolutionized information transmission (communication and interconnects), while electronics, in parallel, has dominated information transformation (computation). This leads naturally to the following question: how can we unify the advantages of the two as effectively as possible? [17]. CMOS gates only draw energy from the rail when and where called upon; however, the energy required to driving an interconnect from one gate to the next dominates CMOS circuit energy use. Relaying a signal from gate to gate, especially using a clocked scheme, induces penalties in latency and bandwidth compared to an optical waveguide passively carrying multiplexed signals.

This suggests that starting up a new architecture from a photonic interconnection fabric supporting nonlinear opto-electronic devices can be uniquely advantageous in terms of energy efficiency, bandwidth, and latency, sidestepping many of the fundamental tradeoffs in digital and analog electronics. It may be one of the few practical ways to achieve ultrafast, complex on-chip processing without consuming impractical amounts of power [18].

Fig. 1: Speed and efficiency metrics that are accessible by various neuromorphic hardware platforms. On the top-right: the photonic neuron platforms studied in Ref. [18]. Hybrid III-V/Si stands for III-V/Silicon hybrid platform spiking neural network photonic integrated circuit. Sub-*l* stands for sub-wavelength photonics. The other points refer to recent electronic neuromorphic hardware, discussed in

2. R. Dennard, F. Gaensslen, W.-N. Yu, L. Rideout, E. Bassous, and A. Le Blanc, "Design of Ion-Implanted Small MOSFET ' S Dimensions with Very," IEEE Journal of Solid State Circuits **9**, 257–268 (1974).

3. H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark Silicon and the End of Multicore Scaling," IEEE Micro **32**, 122–134 (2012).

4. D. A. B. Miller, "Attojoule Optoelectronics for Low-Energy Information Processing and Communications: a Tutorial Review," (2016).

5. M. B. Taylor, "Is dark silicon useful? Harnessing the four horsemen of the coming dark silicon apocalypse," Proc. of Design Automation Conference pp. 1131–1136 (2012).

6. C. Kachris and I. Tomkos, "A Survey on Optical Interconnects for Data Centers," IEEE Communications Surveys & Tutorials