Energy-Efficient and Scalable Bio-inspired Nanophotonic Computing

Extended Abstract

Mohammadamin Nazirzadeh, Pouya Fotouhi, Mohammadsadegh Shamsabardeh, Roberto Proietti, S. J. Ben Yoo
University of California
Davis, California
sbyoo@ucdavis.edu

ABSTRACT

This paper discusses bio-inspired neuromorphic computing utilizing nanophotonic, nanoelectronic, and NEMS technologies integrated into reconfigurable 2D-3D integrated circuits as hierarchical neural networks. The goal is to achieve $\geq$ 1000$\times$ improvements in energy-per-operation compare to the state-of-the-art implementations of neural networks on Von-Neumann based computers. We combine nanophotonic and nanoelectronic technologies to build energy-efficient ($\sim$ 10 fJ/b) artificial spiking neurons with required functionality (spiking, integration, thresholding, reset). Photonic interconnects exploiting 2×2 NEMS-MZIs enables distance independent propagation of signal with weighted addition among the neurons as well as possibility of on-line learning capability. Using low-leakage nanophotonic and nanoelectronic devices, and NEMS, the static power consumption of the system can be decreased down to nearly zero. Realizing 2D-3D phothonic integrated circuit technologies, the proposed system can overcome the scalability limitations of current neuromorphic computing architectures.

KEYWORDS

Neuromorphic Computing, Nanophotonic, Spiking Neurons

1 INTRODUCTION

Large-scale computing is critical for state-of-the-art highly expensive computations. More data intensive applications are emerging and scaling to exascale computing is inevitable. However, the power consumption of an exascale computer is so high that the operation cost will be more than a billion dollars per year. Therefore, extremely low power consumption and highly scalable heterogeneous computing architectures are required for addressing the future computational needs. Bio-inspired neuromorphic hardware systems have been proposed to address these challenges in Von Neumann architectures. Fig. 1 illustrates the possible future computing system combining Von-Neumann and Neuromorphic computing architectures using 3D integrated nanotechnologies to take advantages of both detail-oriented and artificial-intelligence. In this paper we propose an optical neuromorphic architecture using attojoule interconnect solutions exploiting quantum impedance conversion and nanoMEMS latching components, nano-LED or nanolasers, nanophotonic detectors, and 3D integration of nano-circuits. The goal is to achieve femtojoule per bit power consumption, paving the way to solve the current challenges of scaling to exascale computing.

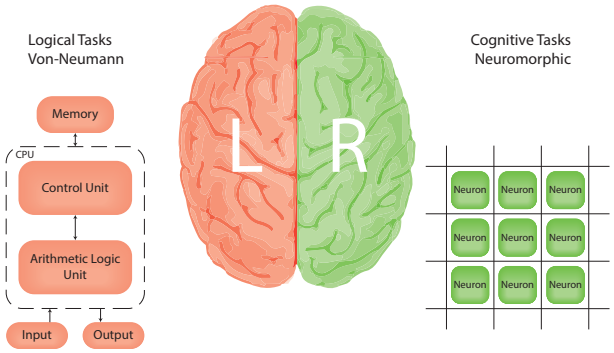

Figure 1: A conceptual illustration of a possible future computing system combining detail-oriented and artificial-intelligence based computing combining Von Neumann and non-Von-Neumann architectures. In both cases, 3D integrated nano technologies will be essential.

2 BACKGROUND

Deep Neural Networks running on Von-Neumann architectures are mastered in cognitive tasks such as the game of Go [1] and face-recognition [2] and surpassed human in performance in some aspects. However, today's supercomputers consume orders of magnitude greater energy than a human brain while processing such tasks. Therefore, more energy-efficient architectures are required to decrease the power consumption. In a human brain, there is approximately 100 Billion neurons, approximately $1,000$ trillion synaptic connections, and a neuron is connected to up to $\sim$ $10,000$ other neurons. Based on [3–5], communication of neurons consume approximately $20,500$ ATP[1]/bit corresponding to 1.04 fJ/bit at 32 bit/s which is extremely efficient. Also, $10^6 - 10^7$ ATP/bit [3] or $50 - 500$ fJ/bit is consumed by a neuron when firing a spike which is proportional to the length of the transmission path, because the axon is a dispersive and lossy medium. Recently, bio-inspired neuromorphic

[1] Adenosine triphosphate (ATP) is a small molecule used in cells for intracellular energy transfer

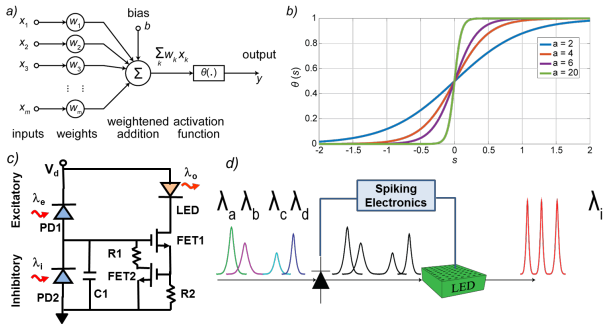

Figure 2: (a) A simple example of a nonlinear model of a neuron. (b) Sigmoid activation function for different slope parameters. (c) A proposed nanophotonic neuron. (d) Illustration of the response encoded in spikes.

computing architectures have been proposed [6, 7]. In particular, IBM has developed TrueNorth based on 28 nm CMOS technology and achieved $176,000\times$ less energy consumption comparing to the state-of-the-art Von-Neumann hardware system [6]. However, these approaches present the following limitations:

- Lack of online training features makes the training process energy and time consuming.
- Long electrical wires bring large capacitance and high interconnect energy consumption. i.e. The TrueNorth chip consumes 2.3 pJ/bit with an additional 3 pJ/bit for every cm transmission.
- Electronic interconnect topologies are typically in four directions (North, East, West, South) and required a number of repeaters.
- 2D and single hierarchical interconnection topology limits their scalability.

The next section discusses a bio-inspired nanophotonic architecture that can tackle some of the above challenges.

3 PROPOSED ARCHITECTURE

In a recent article [8], assuming $\sim$ 19 dB (80×) link loss budget and $\sim$ % wall plug efficiency of the light source, approximately 10 fJ/b interconnect is practically presented. This interconnect solution exploits quantum impedance conversion [9] where signal is transmitted in a close integration with electronics with $\sim$ 1 fF capacitance, instead of charging large interconnect capacitances. We propose then a nanophotonic neuron with 10 fJ/b energy efficiency (Fig. 2). Using nanophotonic repeaters, the proposed nanophotonic neurons can be connected to $10 - 100$ low-loss waveguides providing distance-independent communication among neurons. Reconfigurable photonic interconnects can be formed using nanoMEMS (NEMS) to form hierarchical synaptic interconnects that are able to remember the synaptic weights using NEMS components (Fig. 3). We can conquer the scalability limitation by 2D and 3D photonic integration of the nanophotonic neurons. Silicon CMOS compatible 2D and 3D integrated circuit technologies paves the way to achieve manufacturable energy-efficient computing systems in the future (Fig. 4).

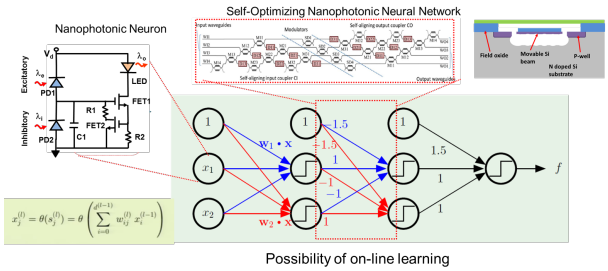

Figure 3: A 3-layer Nanophotonic Neural Network with Nanophotonic Neurons at each node and the self-optimizing nanophotonic neural network with 2x2 NEMS-MZI between each layer.

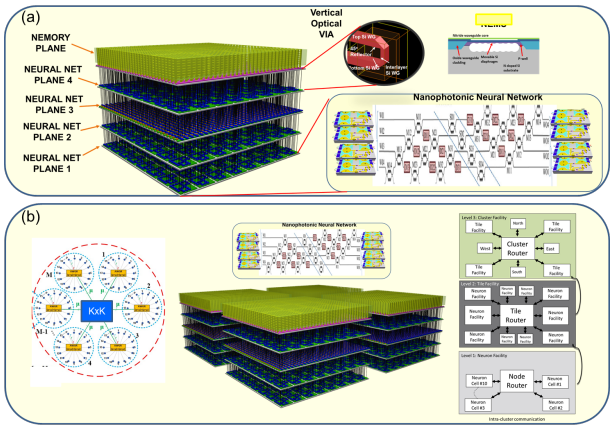

Figure 4: (a) Multi-layer nanophotonic neural network computing platform in a 3D nano-integrated circuit. (b) Hierarchical clustering for scalable computing.

REFERENCES

[1] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[2] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[3] Simon B Laughlin, Rob R de Ruyter van Steveninck, and John C Anderson. The metabolic cost of neural information. *Nature neuroscience*, 1(1), 1998.

[4] Simon B Laughlin. Energy as a constraint on the coding and processing of sensory information. *Current opinion in neurobiology*, 11(4):475–480, 2001.

[5] Simon B Laughlin and Terrence J Sejnowski. Communication in neuronal networks. *Science*, 301(5641):1870–1874, 2003.

[6] Paul A Merolla, John V Arthur, Rodrigo Alvarez-Icaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668–673, 2014.

[7] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana. The spinnaker project. *Proceedings of the IEEE*, 102(5):652–665, May 2014.

[8] David AB Miller. Attojoule optoelectronics for low-energy information processing and communications. *Journal of Lightwave Technology*, 35(3):346–396, 2017.

[9] David AB Miller. Optics for low-energy communication inside digital processors: quantum detectors, sources, and modulators as efficient impedance converters. *Optics Letters*, 14(2):146–148, 1989.