Energy-Efficient and Scalable Bio-inspired Nanophotonic Computing

Mohammadamin Nazirzadeh, Pouya Fotouhi, Mohammadsadegh Shamsabardeh, Roberto Proietti, S. J. Ben Yoo

Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616, USA

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. We will pursue extremely energy-efficient computing with greater than 1000x improvements in energy-per-operation compare to the state-of-the-art implementations of neural networks on Von-Neumann based computers.

Motivation

Large-scale computing platforms, also called warehouse-scale computing, have significantly transformed our lives over the past decades. Emerging applications are even more data intensive. However, scaling to an exascale system for both floating-point (e.g. for climate modeling) and fixed-point (e.g. deep neural network for pattern recognition) are facing severe challenges since the Dennard's power-scaling law failed to keep up with the Moore's law in 2006 [1].

Bio-inspired neuromorphic computing architectures have been proposed. In particular, IBM has developed TrueNorth based on 28 nm CMOS technology and achieved 176, 000× less energy consumption comparing to the state-of-the-art Von-Neumann hardware system [2]. 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

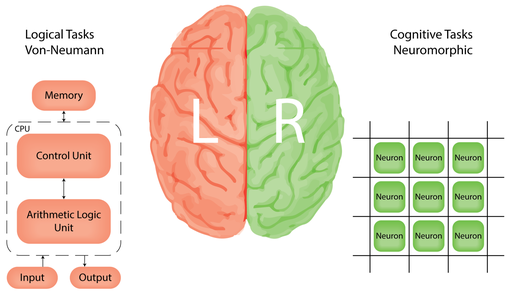
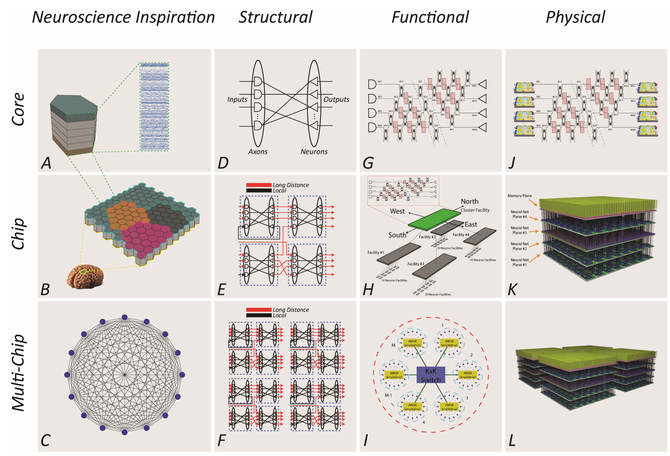


Figure 1. A conceptual illustration of a possible future computing system combining detail-oriented and artificial-intelligence based computing combining von Neumann and non-von-Neumann architectures. In both cases, 3D integrated nano technologies will be essential.

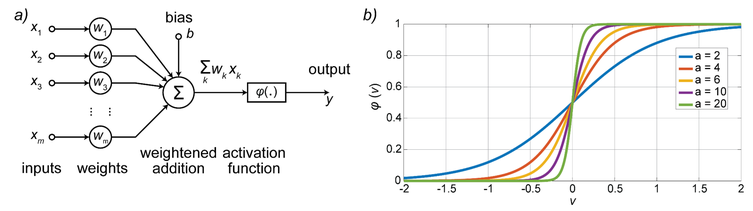
[1] [2]

Overview of the Proposed Approach

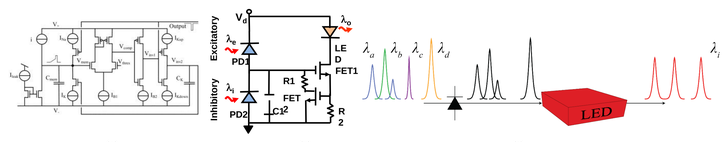


Proposed 3D Neuromorphic Nanocomputing architecture. (A) A canonical cortical microcircuit. (B) The cortex's two-dimensional sheet. (C) The long range connections between cortical regions. (D) Structure of a neurosynaptic core with axons as inputs, neurons as outputs. Multicore networks at (E) chipscale and (F) multichip scale are both created by connecting a neuron on any core to an axon on any core.(G) Functional view of core as a crossbar where horizontal lines are axons, crosspoints are individually programmable synapses. (H) Functional chip architecture is a two dimensional array of cores where long-range connections are implemented by sending spike events (packets) over a mesh routing network to activate a target axon. (I) Hieararchical interconnection network. (J) Optically interconnected neurons. (K) 3D integrated neuromorphic system. (L) multi-3D chip neural networks.

Nanophotonic Spiking Neural Networks



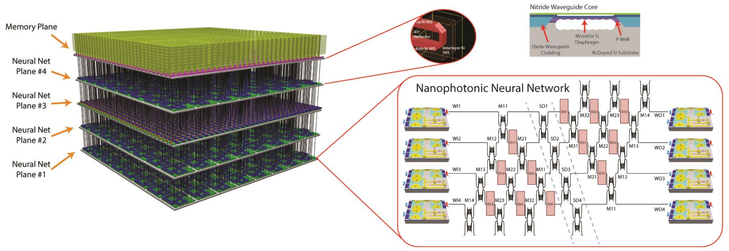
(a) A simple example of a nonlinear model of a neuron which includes (i) synapses, (ii) weighted addition, and (iii) nonlinear activation function. (b) A nonlinear activation function (e.g. sigmoid function) for five different slope parameter .



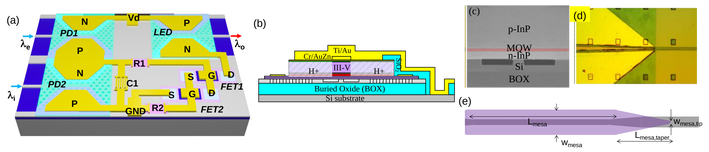
(a) An example circuit designed for generating bio-inspired spike signals [3], (b) a proposed nanophotonic neuron with the spiking electronic circuit, nanophotonic detectors (PD1 for excitatory and PD2 for inhibitory), a nanophotonic LED, and other elements such as resistors and capacitance (C1 inclusive of the capacitance of PDs and FETs) can be tuned to achieve desired sigmoid response function including temporal and rate coding, (c) Illustration of the response that achieves the desired integration and all-or-nothing response encoded in spikes at the output once the photodetector and the LED are connected to the circuit of (b).

[3]

Physical Implementation of Neurons, Synapses and Axons plus Bench Marking

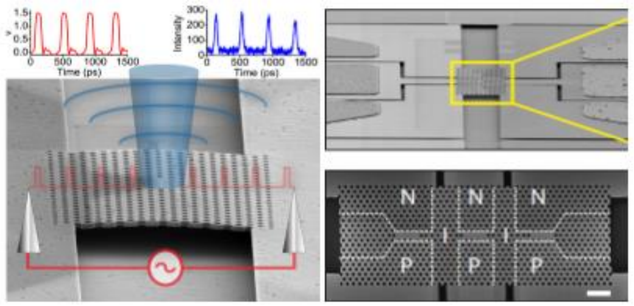


Proposed multi-layer nanophotonic neural network computing platform in monolithic 3D integrated circuit.

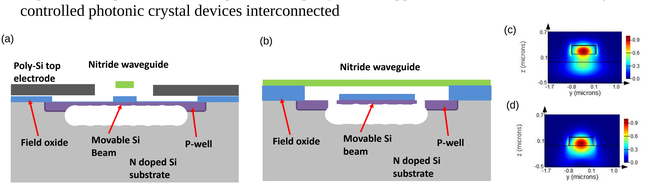


(a) Proposed optoelectronic neuron structure (b) schematic including two InGaAs photonic crystal enhanced photodiodes for excitatory and inhibitory inputs and two FETs on SOI for thresholding and spiking signal generation. A photonic crystal cavity LED for in-plane emission is also incorporate serving as axon function. (b) The nanophotonic photodetectors and LED will be fabricated on silicon utilizing heterogeneous integration by wafer bonding to realize hybrid III-V/silicon nanophotonic devices. Photonic crystal structures, as well as FET devices will be fabricated on silicon-on-insulators while wafer bonding of III-V materials will allow realization of the nanophotonic LED and the nanophotonic detectors. (c) is a cross-section photograph, (d) top view photograph, and (e) a top-view schematic of the hybrid InP Multi-Quantum-Well / silicon semiconductor optical amplifier demonstrated in [4] utilizing a similar fabrication process.

[4]

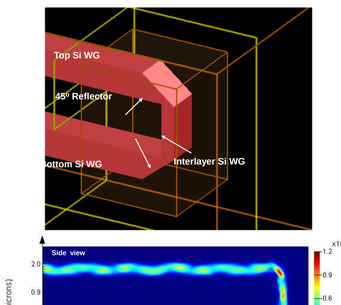


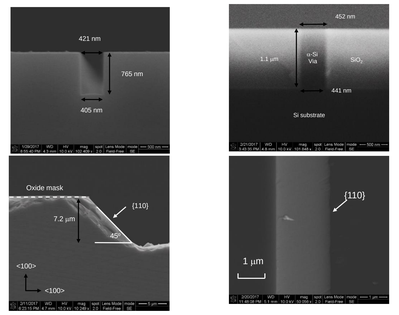
Left: Purcell enhanced single mode LED based on photonic crystal cavity with an embedded lateral pin junction, achieving direct modulation at the 10GHz speed and with <1fJ/bit energy required for operation. [5] Right: lateral pin junction approach with several electrically-controlled photonic crystal devices interconnected



Cross section of phase shift region seen transversally (a) and longitudinally (b) to the waveguide. The movable Si beam has a section of compressively-stressed oxide that makes it buckle down until it is pulled up into close proximity of the waveguide by applying a voltage between the beam and the poly-Si top electrode. The Si beam is P doped to allow a voltage to be applied between it and the N-doped substrate to pull the beam back into its lower stable state. The elongation of the beam caused by the oxide make both the down and up positions stable without the need for holding voltages. Mode profile at 1.55 um wavelength of guided wave with the Si beam in the down position (c) and up position (d). In the down position, the velocity is 1.6·108 m/s and in the up position it is 1.5·108 m/s.

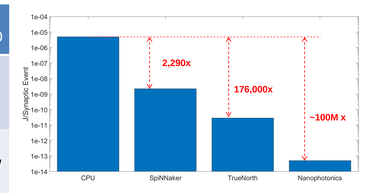
[5]





Vertical Photonic Via designed for neural network plane-to-plane vertical interconnections.

	Conventional Computing System	SpiNNaker	IBM TrueNorth	Nanophotonic (Our Approach)
Energy per Synapsis event (20 Hz firing rate)	>2.6 μJ	20 nJ	26 pJ	≤ 16 pJ
Number of FLOPS or SOPS	4.5 B FLOPS/W	NA	≥46B SOPS/W	≥ 75 SOPS/W



(Left) pJ/Synaptic event and SOPS. (Right) System energy efficiency in pJ per Synaptic event for different neuromorphic approaches. Under a case of 20 Hz average firing rate and 128 active synapses per neuron, the total measured power for TrueNorth is 26 pJ per synapsis event. Measured data shows several pJ/bit for the long distance inter-core communications. As our proposal is targeting fJ/b link nearly independent on the distance, we could estimate that by mapping the TrueNorth architecture on our system, we could potentially get approximately a factor of > 1000x reduction in communication energy and ~500x improvement compared to the state-of-the-art TrueNorth when using a cluster of 64 neurons interconnected through our NxN MZI crossbar.