Silicon Photonics Codesign for Deep Learning

Article *in* Proceedings of the IEEE · February 2020

DOI: 10.1109/JPROC.2020.2968184

CITATIONS
0

READS
920

6 authors, including:

Qixiang Cheng
University of Cambridge
83 PUBLICATIONS **718** CITATIONS
SEE PROFILE

Madeleine Glick
The University of Arizona
159 PUBLICATIONS **1,470** CITATIONS
SEE PROFILE

Meisam Bahadori
University of Illinois, Urbana-Champaign
57 PUBLICATIONS **524** CITATIONS
SEE PROFILE

Keren Bergman
Columbia University
592 PUBLICATIONS **10,057** CITATIONS
SEE PROFILE

INVITED PAPER

Silicon Photonics Codesign for Deep Learning

By Qixiang Cheng, *Member IEEE*, Jihye Kwon, *Student Member IEEE*, Madeleine Glick, *Senior Member IEEE*, Meisam Bahadori, Luca P. Carloni, *Fellow IEEE*, and Keren Bergman, *Fellow IEEE*

ABSTRACT | Deep learning is revolutionizing many aspects of our society, addressing a wide variety of decision-making tasks, from image classification to autonomous vehicle control. Matrix multiplication is an essential and computationally intensive step of deep-learning calculations. The computational complexity of deep neural networks requires dedicated hardware accelerators for additional processing throughput and improved energy efficiency in order to enable scaling to larger networks in the upcoming applications. Silicon photonics is a promising platform for hardware acceleration due to recent advances in CMOS-compatible manufacturing capabilities, which enable efficient exploitation of the inherent parallelism of optics. This article provides a detailed description of recent implementations in the relatively new and promising platform of silicon photonics for deep learning. Opportunities for multiwavelength microring silicon photonic architectures codesigned with field-programmable gate array (FPGA) for pre- and postprocessing are presented. The detailed analysis of a silicon photonic integrated circuit shows that a codesigned implementation based on the decomposition of large matrix–vector multiplication into smaller instances and the use of nonnegative weights could significantly simplify the photonic implementation of the matrix multiplier and allow increased scalability. We conclude this article by presenting an overview and a detailed analysis of design parameters. Insights for ways forward are explored.

KEYWORDS | Deep learning; microring resonator (MRR); neural network; photonic integrated circuit (PIC); silicon photonics.

I. INTRODUCTION

Deep learning is an extraordinarily popular machine-learning technique that is revolutionizing many aspects of our society. Machine learning addresses a wide variety of decision-making tasks such as image classification [1], audio recognition [2], autonomous vehicle control [3], and cancer detection [4]. Matrix multiplication is an essential but time-consuming operation in deep learning computations. It is the most time-intensive step in both feedforward and backpropagation stages of deep neural networks (DNNs) during the training and inference and dominates the computation time and energy for many workloads [1]–[3], [5], [6]. Deep learning uses models that are trained using large sets of data and neural networks with many layers. Since DNNs have high computational complexity, recent years have seen many efforts to go beyond general-purpose processors and toward dedicated accelerators that provide superior processing throughput and improved energy efficiency.

It has been known for quite a while that matrix–vector multiplication can be performed by optical components taking advantage of the natural parallelism of optics to reduce computation time from $O(N^2)$ to $O(1)$ [7], [8]. Implementing these optical matrix–vector multipliers (OMMs), however, required the use of bulky inefficient optical devices. In the last several years, the field of silicon photonics has made major progress toward meeting the massive needs of data centers and cloud computing. With silicon photonics, optical components and photonic

Manuscript received March 19, 2019; revised August 19, 2019; accepted January 10, 2020. This work was supported in part by the National Science Foundation (NSF) under Grant CCF-1640108 and in part by the Semiconductor Research Corporation (SRC) under Grant SRS 2016-EP-2693-A. *(Corresponding author: Qixiang Cheng.)*

Qixiang Cheng was with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA. He is now with the Electrical Engineering Division, Department of Engineering, University of Cambridge, Cambridge CB3 0FA, U.K. (e-mail: qc223@cam.ac.uk).

Madeleine Glick, Meisam Bahadori, and **Keren Bergman** are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA.

Jihye Kwon and **Luca P. Carloni** are with the Department of Computer Science, Columbia University, New York, NY 10027 USA.

Digital Object Identifier 10.1109/JPROC.2020.2968184

integrated circuits (PICs) are fabricated leveraging CMOS-compatible silicon manufacturing techniques to enable small-footprint, low-cost, power-efficient data transfers.

OMMs based on silicon photonics represent a promising approach to address the challenge of compute-intensive multiplication in DNNs. An optimal solution must take into account the advantages and drawbacks of the silicon photonic technology along with the requirements of the application. Silicon photonics offers excellent codesign capabilities with off-chip control implemented by field-programmable gate arrays (FPGAs) to achieve accelerated computational gains. To analyze these capabilities in detail, we present the codesign of a DNN in conjunction with the OMM, developing an optical-electrical codesign infrastructure using FPGA control. The FPGA is used for: 1) pre/postprocessing and 2) photonic device control. We identify opportunities for OMM architectures based on multiwavelength silicon microring resonators (MRRs). We analyze and generalize the metrics of the microrings for linearity and reduced sensitivity to perturbations. The OMM can be used in the case of time-consuming, computationally expensive matrix multiplication. In the case of DNNs that are too large to be processed on a single optical chip, we explore methods to share the computation, by using the parallelism at the system level to enable scaling to very large neural networks. In addition, we show how DNNs based on nonnegative weights significantly simplify the photonic implementation of the matrix multiplier and allow increased scalability.

The remainder of this article is organized as follows. In Section II, we present a brief background of advances in deep learning and in silicon photonics. In Section III, we provide an overview of and discuss tradeoffs in the state-of-the-art research in the implementation of silicon photonics for deep learning. Based on the above-mentioned analysis, in Section IV, we propose a codesigned system for deep learning. We first present a detailed analysis of the design parameters and metrics for a silicon PIC that implements an optical matrix multiplier. We generalize the role and characteristics of the silicon microrings, analyzing their limitations (including thermal sensitivities) in order to explore opportunities for optimized OMM structures. We then discuss system-level approaches toward electronic/photonic codesign for improved performance. At the end of this section, we provide insights into future directions and opportunities based on our analysis and the current state-of-the-art and application requirements. Section V concludes this article.

II. BACKGROUND

A. Deep Learning

The fundamental concept of machine learning is that the core computation algorithm is not fully provided by a programmer, but automatically generated or improved by a computer system through experience [9]. The learning system explores a given class of computation models to determine the most suitable model among them based on the training data. One of the model classes that has gained widespread popularity is the DNN, which is the artificial neural network (ANN) with many layers in the network [10]. Inspired by the human brain, the concept of the ANN was first proposed in the 1940s [11]. More recently, with the increased volume of data, computing capability, and research interest, numerous ANNs have shown outstanding performance in machine-learning tasks across various application domains [1]–[4]. Deep learning refers to machine learning using deep ANNs, also called DNNs. Two fundamental classes of ANNs are multilayer perceptrons (MLPs) and convolutional neural networks (CNNs).

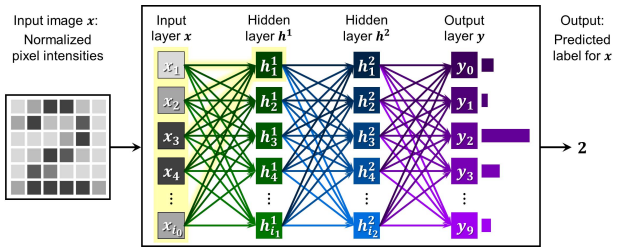

Fig. 1. *MLP for handwritten digit classification. The network consists of four layers (the input layers, two hidden layers, and the output layer) where each layer contains a number of nodes (also called neurons).*

MLPs, also known as fully connected networks (FCNs), are the quintessential DNNs [10]. An MLP represents a function defined by a network consisting of multiple layers of nodes, which are also called neurons or perceptrons. For example, Fig. 1 shows an MLP for the task of recognizing a handwritten digit. The input image is represented as an array of pixel intensity values which are often normalized. The neural network behaves as a function that maps the input image to the probability score for each of the ten digits $(0, 1, 2, \ldots, 9)$. Let i_0 denote the number of pixels in the input image. Then, for an input array $x \in \mathbb{R}^{i_o}$, the neural network (shown in the box in Fig. 1) outputs $y(x) \in \mathbb{R}^{10}$, as follows. The input layer x contains i_0 nodes $x_1, x_2, \ldots, x_{i_0}$. This layer is fully connected to the first hidden layer h^1, which contains i_1 nodes; each node $h^1_k (1 \le k \le i_1)$ is computed as

$$h^1_k(x_1, x_2, \ldots, x_{i_0}) = \mathrm{Act}\left(g^1_k(x_1, x_2, \ldots, x_{i_0}) + b^1_k\right) \quad (1)$$

$$g^1_k(x_1, x_2, \ldots, x_{i_0}) = w^1_{k,1} \cdot x_1 + w^1_{k,2} \cdot x_2 + \cdots + w^1_{k,i_0} \cdot x_{i_0} \quad (2)$$

where Act() denotes an element-wise nonlinear activation function (e.g., ReLU, sigmoid, softmax, tan h), $b^1_k \in \mathbb{R}$ is a bias, and $w^1_{k,j} (1 \le j \le i_0)$ represents the weight of the connection between node x_j and h^1_k (see Fig. 2). Each hidden layer is fully connected to the next layer, and the last layer in the network is the output layer containing ten nodes. The softmax function is often used for nonlinear

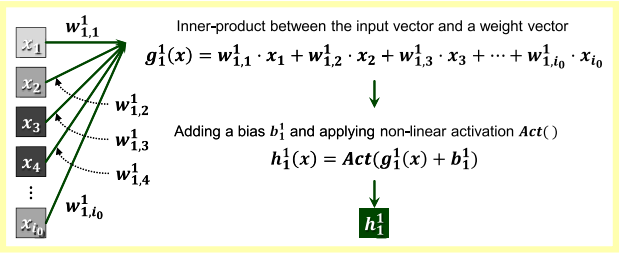

Fig. 2. *Computation for a single node (the first node in the layer h^1) in MLP.*

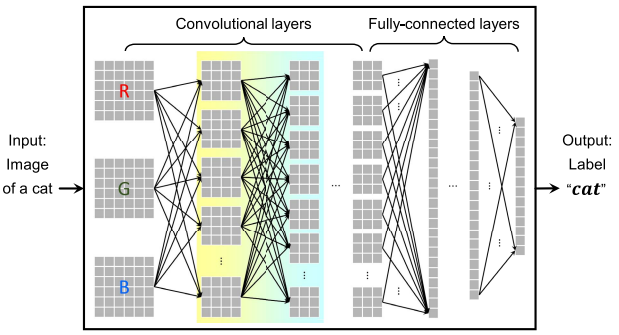

Fig. 3. *CNN for image classification, consisting of convolutional layers followed by fully connected layers. Convolutional layers are elaborated in Fig. 4, and the computation for fully connected layers is depicted in Fig. 2.*

activation of the output layer since it can be interpreted as a probability distribution.

The process of computing the output of a neural network as described above is called *feedforward propagation*. The information stored in the input layer propagates toward the output layer. How it propagates depends on the neural network structure, weights, biases, and activation functions. During the training phase of supervised machine learning, given a large number of (input–output) instances, the values of weights and biases are updated through the gradient descent method, also called *back-propagation* [12]. Then, in the inference phase, a trained network is used to predict the output for a new input instance. With this approach, MLPs were among the first and most successful nonlinear learning algorithms [10]. The nonlinear activation plays a key role in ANNs. Without the ANN, the function expressed by an MLP is a composition of linear functions (which is linear). By inserting the nonlinear activation, such as ReLU or tan h, the resulting function becomes a composition of nonlinear functions, which can express much more complicated concepts.

In addition, the universal approximation theorem states that any continuous function defined on a compact set can be approximated by an MLP with a single hidden layer [13], [14]. Nevertheless, it does not address how many nodes are required in the hidden layer or how to learn the weights and biases of such an MLP. Empirically, the accuracy of the trained networks improves as the number of nodes per layer increases, and as the number of layers increases. This motivated the advancement of DNNs.

CNNs were first proposed by LeCun *et al.* [15] for handwritten digit recognition, and they have outperformed many proposed MLPs, especially for more complex tasks such as colored image classification. Fig. 3 illustrates the overview of a CNN for image classification. The input image is stored across three channels, each representing the red, green, or blue intensities. As shown in Fig. 4(a), a convolutional layer (layer $L+1$) usually contains multiple channels, and the values of nodes in each channel are computed using the information from all channels in the previous layer (layer L). Fig. 4(b) gives a closer look at the connection between an input channel from the previous layer (channel A in layer L) and an output channel in a convolutional layer (channel B in layer $L+1$). A convolution kernel (of size 3×3 in the example) dedicated to this connection defines how to obtain a value for each node in the output channel from a small neighbor [of size 3×3 in the input channel; see Fig. 4(c)]. The kernel slides both vertically and horizontally on the input channel to cover all nodes in the channel, and the convolution result is propagated to the node in the associated position in the output channel. The amount by which the kernel slides is called the stride and this is often set to 1. Around the boundary of the input channel, additional nodes of the value zero can be padded before the convolution. When the kernel is of size $R \times R$, a padding of size $\lceil R/2 \rceil$ is commonly applied. At each node in the output channel, bias and activation function are applied to the summation of the corresponding convolution results. To summarize, the value of a node $z_{c,d}^{L+1,B}$ at (c, d)-coordinate on channel

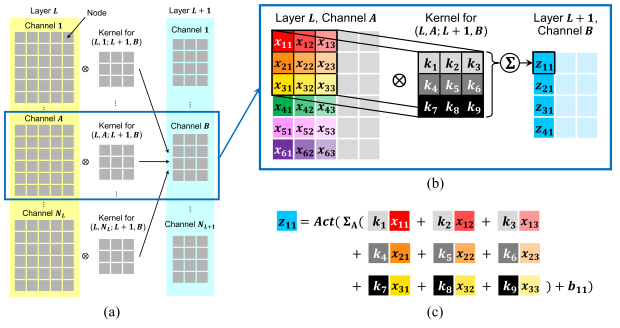

Fig. 4. *Overview of convolutional layers. A convolutional layer consists of one or more channels where each channel contains a number of nodes. (a) Every channel in the previous layer is connected to each channel in the next layer. (b) Connection between one input channel (in layer L) and one output channel (in layer L+1). The convolution between a set (in the black square) of nodes in the input channel and the convolution kernel ($k_1, \ldots, k_9$) contributes to one node in the output channel. (c) Computation for a single node in the output channel. Σ_A denotes the summation over all input channels A.*

B in layer $L+1$ is computed as

$$z_{c,d}^{L+1,B}(z^{L,1}, z^{L,2}, \ldots z^{L,N_L};\, k^{L,1;L+1,B}, k^{L,2;L+1,B}, \ldots, k^{L,N_L;L+1,B}) = \mathrm{Act}\left(\Sigma_{A=1}^{N_L} v_{c,d}^{L+1,B}(z^{L,A};\, k^{L,A;L+1,B}) + b_{c,d}^{L+1,B}\right) \quad (3)$$

$$v_{c,d}^{L+1,B}(z^{L,A};\, k^{L,A;L+1,B}) = \Sigma_{\alpha=1}^{M}\Sigma_{\beta=1}^{M}\left(k_{\alpha,\beta}^{L,A;L+1,B} \cdot z_{c-\left\lfloor \frac{M}{2} \right\rfloor+\alpha,d-\left\lfloor \frac{M}{2} \right\rfloor+\beta}^{L,A}\right) \quad (4)$$

where Act() is an activation function, $b_{c,d}^{L+1,B} \in \mathbb{R}$ is a bias associated with this output node, $z^{L,A}$ denotes channel A in the previous layer L, N_L represents the number of channels in layer L, and $k^{L,A;L+1,B}$ refers to the convolution kernel of size $M \times M$ defined for the connection between the channel A in layer L and the channel B in layer $L+1$.

Optionally, a convolutional layer may be followed by a pooling layer that reduces the size of the representation by pooling neighbors of $R \times R$ nodes, where R often takes a small value such as 2, 3, 4, or 5. Most commonly used pooling functions are *maximum* (i.e., taking the maximum value from the $R \times R$ neighborhood), *average, median*, and *stochastic*.

The network size and computational complexity of state-of-the-art DNNs have generally increased over the decades. Meanwhile, much research has also been conducted on the accelerated and efficient computation of DNNs [16]. For both MLPs and CNNs, the core computation requirements during feedforward propagation are inner products of two vectors, or matrix–vector multiplications [5], [6], [17]. Both the weight product function $g()$ for fully connected layers [in (2)] and the convolution function $v()$ for convolutional layers [in (4)] can be naturally translated into vector–vector or matrix–vector multiplications. Graphical processing units (GPUs) have been extensively exploited to accelerate this type of computation, mainly leveraging their inherent feature of single-instruction–multiple-data parallelism [18], [19]. In addition, there has been growing interest in designing custom hardware accelerators and reconfiguring the DNNs for higher efficiency [20]. Haensch *et al.* [5] have proposed in-memory analog computation for DNNs and have analyzed nonvolatile memory material candidates. Amiri *et al.* [21] have proposed a multiprecision CNN framework on an FPGA-CPU heterogeneous device.

B. Silicon Photonics

Although GPUs, FPGAs, and application-specific integrated circuits (ASICs) have received extensive interest for developing dedicated hardware accelerators in deep-learning calculations [22]–[24], photonics has long been recognized as a promising alternative to address the fan-in and fan-out problems for linear algebra processors [25], [26]. A few unparalleled features motivate the exploration of a photonic implementation.

1) The power consumption for data transfer that accounts for a large portion in electronic ASICs [16] can be greatly reduced by leveraging state-of-the-art optical transceivers. In addition, once a neural network is trained, the matrix configuration can be passive and optical signals can be processed with no additional power consumption [27].
2) The operation bandwidth of such an OMM could potentially match that of the photodetection rate (typically in 100 GHz), which can be at least over an order of magnitude faster than the electronic system (typically restricted to the clock rate of a few GHz).
3) The OMM could have significantly lower latency, since the electronic hardware accelerators still rely on electronic transport that is bounded by the speed and power limits due to RC parasitic effects. Early demonstrations of photonic solutions were implemented with bulky free-space optics [25], [26], which required rigorous calibration for phase matching and have extreme scaling difficulties. Current photonic integration platforms provide opportunities for highly scalable solutions that improve energy efficiency and significantly reduce overhead of assembly, calibration, synchronization, and management [28].

Over the last two decades, silicon has been shown to be an excellent material platform for fabricating photonic devices, and processes have been developed to permit the reuse of CMOS manufacturing infrastructure to build complex PICs. It is, therefore, not surprising that silicon photonics is now widely accepted as a key technology in next-generation communications systems and data interconnects [29]. On the one hand, following the example of the electronic fabless semiconductor industry, process design kit (PDK) libraries are being developed and standardization is being encouraged by the silicon photonics industry and users for broader accessibility [30], [31]. On the other hand, component customization is driven by a number of research groups and companies that design a large variety of specialized photonic components [32]–[34]. The ability to include increasing numbers of a wide range of optical components at the wafer scale has led to a powerful class of silicon-based PICs [35]. Such integration technology fundamentally improves circuit-level performance by reducing the complexity in assembly, calibration, and synchronization. As it matures, sustained increases in the functionality, performance, and reliability of circuits are enabled. This, in turn, stimulates new research directions leveraging the large-scale photonic integration capabilities [27], [36]–[40]. Lightwave signals have been manipulated in their intensity and phase at the space, wavelength, polarization, and mode dimensions, for data transmission [33], [41], switching [42]–[44], and processing [27], [37], [40], in both digital [33], [41] and analog formats [27], [39], [40].

In addition, in recent years the ecosystem of silicon photonics has been extended to enable further functionality. The ability to add CMOS-compatible materials, such as germanium (Ge), Ge-rich GeSi, and silicon nitride (SiN), to the silicon-on-insulator (SOI) platform has significantly

enriched the component library and enhanced circuit-level performance. Notable examples include the Ge-on-Si photodiodes (PDs) [45], high-speed GeSi modulators [46], and the ultralow loss Si/SiN multilayer structure [47]. The development of heterogeneous integration [48], [49] as well as breakthroughs on the direct growth of III–V quantum dot materials on silicon substrates [50] further completes the ecosystem, enabling a *System-on-Chip*.

The Mach–Zehnder interferometer (MZI) and the MRR are two of the most common functional building blocks in many photonic systems, such as modulators [32], [51], [52], filters [34], [53], multiplexers [54], [55], switches [56]–[58], and computing systems [27], [59], [60]. The MZI was first proposed over a century ago to determine the relative phase shift variations between two collimated beams derived by splitting the light from a single source. Later work extended this concept to manipulate the probability of light arriving at either port, by precisely controlling the phase difference between the two arms [61]. Integrated MZIs generally consist of two 3-dB couplers with phase shifters embedded in each of the two arms. Detailed design considerations can be found in [52], [53], and [62]. An MRR consists of an optical waveguide which is looped back on itself and coupled waveguides. Resonance occurs when the optical path length of the resonator is exactly a whole number of wavelengths and thus multiple resonances are supported. The spacing between these resonances is called the free spectral range (FSR). Similarly, a phase shifter can be embedded in the resonator to tune the optical path length in order to shift the resonance spectrum. The properties of MRRs are extensively described in the literature [63], [64], as well as their design considerations, performance metrics, and potential challenges [29], [32], [34], [54], [63], [64]. We discuss the applications of the MRR in more detail below.

III. SILICON PHOTONICS FOR DEEP LEARNING

This emerging area of research has been stimulated by recent results in which silicon photonics has been utilized to implement optical neural networks based on a spatial multiplexing technique with coherent interference [27] and a spectral multiplexing technique with wavelength filters [60]. In this section, we give a detailed overview of this recent progress in programmable silicon photonics for deep-learning hardware accelerators.

A. Linear MZI-Based Meshing Optics With Orthogonal Spatial Modes

Pioneered by the work of Reck *et al.* [65] showing that a mesh of 2×2 beam splitters and phase shifters in the form of an MZI can be programed to enable independent control of amplitude and phase of light for a set of optical channels, various novel architectures and design principles based on a cascade of MZIs have been proposed and demonstrated for both classical and quantum applications [27], [37], [39], [66]–[68]. These works are also referred to as "programmable linear optic processors" [69].

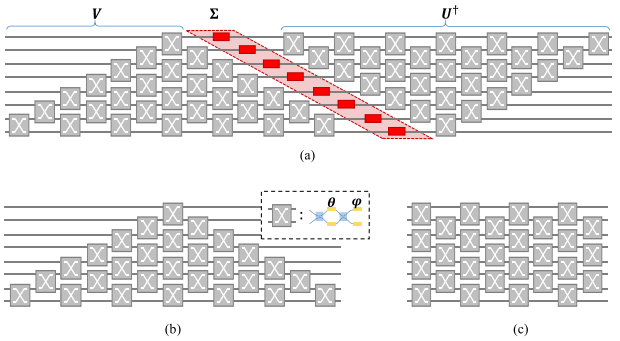

Fig. 5. *(a) Universal linear mesh network comprising two unitary matrices and a diagonal matrix to set amplitude and phase, as proposed in [68]. Universal unitary matrix proposed (b) by [65] and (c) by [66].*

Phase shifters that are embedded in the arms of MZI units are used to control the interference of beams at the combining stage, while a pair of external phase shifters is employed in order to set a differential output phase. This allows the control of relative amplitude and phase of the beams at each stage and thus the programing of the mesh. With specific interconnection patterns, universal linear optical components can be obtained [66]–[68], [70], [71].

Whereas most of the mesh networks are explored as universal linear optics for unitary operations [37], [65]–[67], [70], [71], *Miller* proposed a design method that implements arbitrary, nonunitary matrices, as shown in Fig. 5(a) [68]. This approach describes a self-configuring universal linear mesh that employs a set of orthogonal beams. The mathematics behind this design demonstrates that any linear optical device can be factorized using the singular value decomposition (SVD), as $D = V\Sigma U^{\dagger}$, where V and $U^{\dagger}$ are unitary matrices and Σ is the diagonal matrix [68]. Theoretically, the universal unitary matrices of V and $U^{\dagger}$ can be implemented following the designs proposed by Reck *et al.* [65] [see Fig. 5(b)] and Clements *et al.* [66] [see Fig. 5(c)], and the diagonal matrix Σ can be represented by an array of modulators that can set amplitude and phase [68], as illustrated by Fig. 5(a). The unitary matrices of V and $U^{\dagger}$ can be further decomposed to analytically define the values of beam splitters, i.e., phase settings of MZIs [65], [66].

The recent work by Shen *et al.* [27] proposed a novel architecture (see Fig. 6) for an optical neural network that offers hardware acceleration for deep-learning applications. Vectors were encoded in the intensity and phase of light and then fed into each layer of the network, which comprised an optical interference unit (OIU) and an optical nonlinearity unit (ONU). Although the ONU function was based on a computer to act as a saturable absorber, the OIU was implemented using a silicon PIC to perform the optical matrix multiplications following Miller's design, which leverages the SVD [68]. This optical device consists of 56 programmable MZI units, each of which has two 50:50 power splitters and two pairs of phase shifters parameterized by (θ, φ). The power splitters/combiners are

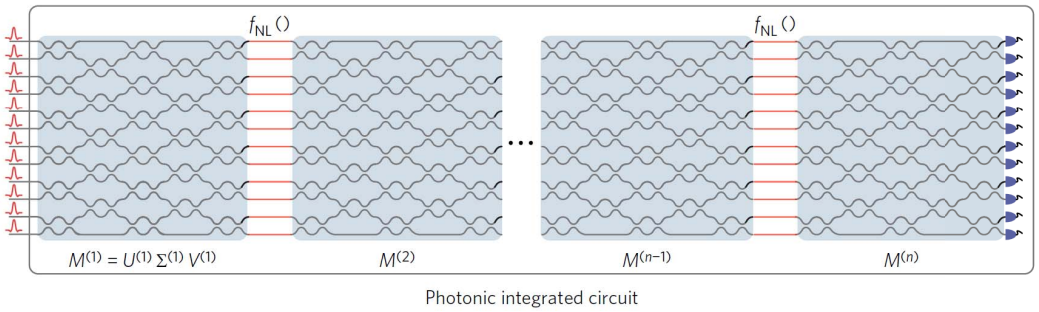

Fig. 6. ***All-optical architecture for integrated neural network [27].***

realized by directional couplers and the $\pi/2$ phase difference between the two outputs ensures the unitary property of its transformation. As a nonapplication-specific PIC, one matrix transformation requires two passes through the chip for: 1) $V\Sigma$ and 2) $U^{\dagger}$. The required orthogonal beams are implemented by a set of coherent spatial modes. This device does not use on-chip detectors for self-alignment. However, other generic approaches for setting up meshes can be leveraged to enable the calibration of phase disorders due to fabrication variations, such as the one described in [72]. In addition, the broadband nature of MZIs does not have a strong requirement for local phase stabilization, although on-chip thermal crosstalk could be a significant cause of phase errors.

Neural network training algorithms [73] can be leveraged to train the matrix parameters for different layers. Each layer contains a set of weights, which can be decomposed into phase settings and then programed into the OIU. By implementing a two-layer optical neural network with four neurons per layer, a primitive task for vowel recognition was executed and achieved an accuracy of 76.7% [27]. Compared to the accuracy of 91.7% by execution with a conventional 64-bit digital computer, the key limiting factor for the accuracy of the optical neural network can be attributed to the computational resolution. The phase-encoding noise and the photo-detection noise are believed to be the primary factors causing reduced resolution [27]. This is also reflected in the fidelity analysis showing that the percentage error for each output of the SU(4) unitary matrix core is approximately 2.24% [27], which bounds the system's effective resolution. Suppressing on-chip thermal crosstalk and lowering photo-detection noise would thus lead to a superior computational resolution of the network.

The work described above shows an impressive example of applying silicon photonics to deep-learning applications; yet, three factors, in particular, might bound the practicability of this approach.

1) Limited Scalability of Neurons: Let N denote the number of neurons. The optical depth (the number of MZI units traversed through the longest path) for the unitary matrix is given as $2N-3$ and as N in the scheme by Reck *et al.* [65] and by Clements *et al.* [66], respectively. This, therefore, leads to a total optical depth of $2N-1$ (with output reflected for a more compact layout [68]) and of $2N+1$, respectively, for the optical device that implements the arbitrary linear transformation using SVD encoding where the diagonal matrix Σ is implemented by an array of MZIs. Note that although the device using Reck *et al.* [65] design has a slightly smaller optical depth, the Clements *et al.* [66] layout is shown to be more tolerant to component loss in realistic interferometers, maintaining high fidelity. The optical depth increases linearly with the number of neurons (N) by a factor of 2 which directly translates into additional loss in silicon photonics integrations. This additional loss could quickly outpace the optical power link budget and significantly deteriorate the system signal-to-noise ratio, thus limiting the computational resolution.

2) Error Accumulation: Whereas the on-chip thermal crosstalk can be suppressed, the finite encoding precision on phase settings will remain as the fundamental limitation for the optical neural networks with high computational complexity. The phase errors, in particular, accumulate when the lightwave signal traverses the MZI mesh with an optical depth of $2N+1$. In addition, such errors propagate through each layer of the network, which ultimately restricts the depth of the neural network.

3) Complex Encoding Scheme of Matrix: The SVD method provides a perfect solution to decompose an arbitrary linear transformation. However, mapping the trained matrix parameters to the phase settings of the MZI mesh consumes additional computational power.

B. Microring Weight Banks for Spiking Networks

Inspired by the field of neuroscience in which biological neurons communicate by short pulses, spike processing, with this integrate-and-fire neuron model, has been proposed to exploit its massive parallelism potential in computation [74]. The cornerstone of the communication protocol is the spike coding scheme, which is digital in amplitude and analog in pulse timing [75]. Input spikes from multiple sources are multiplied by a set of weight factors and temporally integrated to trigger a neuron firing a single output spike if the threshold is satisfied [76]. It has been recently recognized that photonics can be a powerful alternative to the microelectronic platform to implement such a spike processing system, given the significant

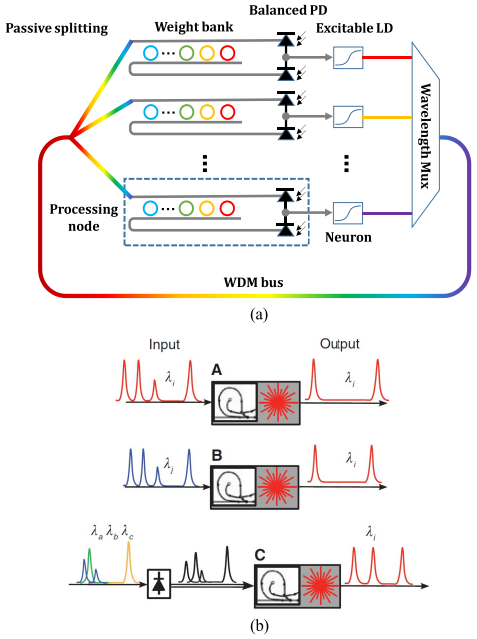

Fig. 7. *(a) Broadcast-and-weight spiking network proposed by [60]. (b) Classification of semiconductor excitable lasers [77].*

advancement in both excitable lasers for the nonlinear processing, and analog PICs for the linear processing [77].

An on-chip optical architecture, named broadcast-and-weight, was proposed by Tait *et al.* [60] to implement scalable photonic spike processing networks to connect parallel neurons. As illustrated in Fig. 7(a), each spiking laser represents a neuron, and the optical neural network connects the output of each neuron to multiple other neurons making use of wavelength division multiplexing (WDM). In contrast to the spatial multiplexing approach, channelization of the spectrum can somewhat simplify the interconnect network of neurons, as WDM channels can coexist in a single bus waveguide channel without interfering. The group of neurons that each utilizes a distinct wavelength share a common bus waveguide, as shown in Fig. 7(a). The broadcast can be simply realized by passively splitting the bus waveguide to connect each of the neurons, enabling the all-to-all connection [60]. Each neuron is attached to a weight processing unit which is used to execute the linear transformation function for the N incident WDM signals that represent N neural nodes including itself. In this case, being capable of independently manipulating each weight is critical for creating differentiation among WDM channels. The silicon add–drop MRR is a natural choice due to its wavelength selective nature, as well as its cascadability, and continuous power-ratio-tunable feature [54]. The bank of cascaded MRRs, as an array of reconfigurable add–drop filters, imprint the weight coefficient to each corresponding channel. In a network of N neurons with N wavelength channels, each neuron incorporates a bank of N MRR filters, leading to a total number of N^2 MRRs. The through port and drop port of the cascaded MRRs are, respectively, connected to create two subsets of weighted power, each connected to one of the balanced PD pair that performs the summation by incoherently aggregating the total incident optical power. The layout of the balanced PD subsequently enables subtraction between the two subsets of weighted powers for inhibitory weighting. The weighted sum is then used to excite a spiking laser neuron and three classifications of semiconductor excitable lasers are shown in Fig. 7(b) [77]. When the temporal integration of weighted pulses can push the gain above the lasing threshold, the neuron releases a spike. Otherwise, the system stays at rest.

As a key constituent element, the MRR weight bank has been carefully studied [78]–[81], since its scalability and tunability are closely tied to the performance limits of the optical neural network. Quantitative analysis was provided to measure the scaling of channel count, N, for an MRR filtering bank, illustrating the limiting factors of interchannel crosstalk, insertion loss, and more importantly, the bus length that causes coherent interactions between adjacent MRRs [78]. Similar to the MRR devices in data communication links, the interchannel crosstalk and cascading loss are the two fundamental constraints for system scale-up [54]. However, in contrast to the (de-)multiplexing-oriented designs that have only one common bus, the bus length becomes a key factor in the weight bank design that brings about multi-MRR coherent interactions due to the two bus configuration. This inevitably introduces another dimension of design complexity. Such interchannel interference also deteriorates the independent control of the WDM channels, as the weights cannot be linearly separated. A more rigorous calibration process can be undertaken to improve these impairments in the WDM channels. Any power leakage or loss can be counter-balanced by adjusting the corresponding MRR coupling ratio. However, the degradation of the MRR weight tuning range eventually becomes irreparable [78]. For a given system error σ, the tuning range is a critical factor that determines the network's computational resolution, as shown below.

A few efforts have been made to optimize the device design and control plane for microring weight banks in silicon photonic integration platforms [79]–[81]. A continuous range of complementary ($\pm$) weighting has been demonstrated and recent work shows an effective weight setting accuracy of 5.1 bits [81], which is defined by $\log_2[(\mu_{\max} - \mu_{\min})/\sigma]$, where $(\mu_{\max} - \mu_{\min})$ is the tuning range and σ is the measured system error. The chip performance in this experiment is facilitated by photoconductive heaters which provide online feedback of photo-induced resistance to estimate the filter transmission. Considering that MRRs are particularly sensitive to thermal drift [82], the real-time feedback control loop, which tracks thermal fluctuations, including ambient temperature change, self-heating effects, and thermal crosstalk, plays a major role in such a multiresonator system. It, therefore, provides superior performance compared to the feedforward control scheme [79], [80], which relies on fixed prebuilt references.

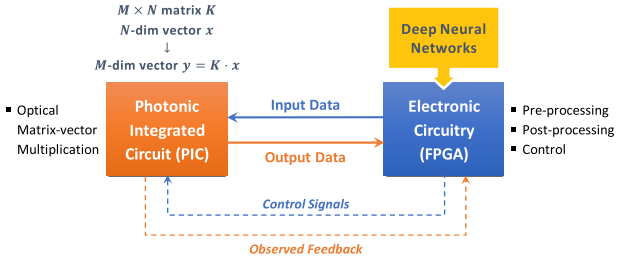

Fig. 8. ***Overview of the proposed codesigned system for deep learning.***

Whereas a set of MRRs sandwiched by two buses that drop power into a balanced photodetector offer complementary (±) weight factors, the closed WDM link makes it difficult to monitor the isolated transmission state for each wavelength channel. Altering the weight factor via shifting the resonance spectrum of individual MRR unit arranged in a cascading scheme would significantly constrain its tuning range, given that all channelized MRR filters coexisting on the same bus have to tightly fit within one FSR. The embedded photoconductive heaters within MRRs provide a limited but adequate solution for neuromorphic applications [77], [81]. However, the adoption of photoconductive effects in the analog computing system may not sufficiently deliver the requirements for optical matrix multiplication with higher resolutions.

C. Discussion

Both of the aforementioned approaches aim at processing an entire ANN application or an entire matrix–vector multiplication on a single optical device. Whereas those approaches may have advantages in the processing speed, the capability of the optical device strictly limits the size of the ANN to be processed. For instance, the optical neural network architecture proposed by Shen *et al.* [27] consists of two layers, each with four neurons, for a primitive machine-learning task of classifying four vowels in speech. However, many machine-learning tasks in practice involve learning more complex functions that take in a large number of inputs. For a handwritten digit recognition task, the number of input neurons is $28 \times 28 = 784$, one for each pixel of the input image, and the number of output neurons is ten, which equals the number of candidate digits [15]. For breast cancer detection, an MLP with 30 input neurons, 500 neurons in each of the three hidden layers, and two output neurons were used to achieve the detection accuracy of 99% [83]. The computation for this MLP includes the multiplication between a matrix of size 500×500 and a vector of dimension 500. It is not feasible or practical to fully optically implement such large neural networks or matrix–vector multiplications using the above approaches due to their limited scalability.

IV. SILICON PHOTONICS CODESIGN FOR DEEP LEARNING

Codesign of silicon photonic and electronic circuits provides new opportunities for efficient computation of deep learning. Silicon photonics has the potential for high-speed analog matrix multiplication. However, the computational requirement for ultra-large DNNs with high accuracy demand may exceed the capability of a single PIC, for high-complexity computing tasks. Our codesign approach, described in this section, explores practical and scalable solutions to process such large neural networks while employing feasible optical devices.

Fig. 8 illustrates an overview of the proposed codesigned system with the electronic circuitry that processes DNNs at the system level, and a PIC that performs optical matrix–vector multiplication of fixed-size inputs. The PIC takes in a matrix K of size $M \times N$ and a vector x of size N, and outputs a vector y of size M such that $y = K \cdot x$.

A. Silicon Photonic OMM

Channelization in the wavelength domain avoids the phase-sensitive designs that require the control of relative phases from different nodes for coherent interference effects. Therefore, the wavelength multiplexing technique provides an elegant solution to address the many-to-one coupling (fan-in), which is a typical problem in neuron networks. Combined with tunable add–drop MRR technology, the direct mapping from the weight matrix to the power coupling ratio of wavelength filters can also eliminate the complex encoding phase. Such simplicity would further boost the validity of optical neural networks as hardware accelerators for deep-learning applications.

The high thermo-optic coefficient of silicon (1.8×10^{-4} K^{-1}) creates a double-edged sword for silicon MRR elements. While it allows effective manipulation of light by the thermo-optic effect, due to its narrow-band nature, this thermal susceptibility can be detrimental to device performance. Therefore, accurate monitoring and control mechanisms are normally required. A number of energy-efficient yet precise locking schemes have been demonstrated [84]–[88] for data communications. The work by Tait *et al.* [81] sheds some light on the feedback control of an analog MRR system, which relies on an estimate of filter transmission. The plasma dispersion effect via either carrier depletion or injection can be leveraged to provide nanosecond-scale tuning mechanism [29]. However, fabrication variations [89], in addition to the self-heating effect and ambient temperature change [82], most often require an additional thermo-optic phase shifter. The electro-optic tuning mechanism also requires attention to the induced electroabsorption loss that compromises the extinction ratio of resonance, thus the resolution of computation. This additional loss disturbs the balance between the coupling power and the round trip loss in the ring cavity from the critical coupling point. The operation condition for critical coupling is discussed in Section IV-A1.

In general, the on-chip thermal crosstalk is a primary culprit of the system instability for MRR-based silicon photonic circuits. Hereby, we start with an analytical model of add–drop MRRs to provide an insight into constraints

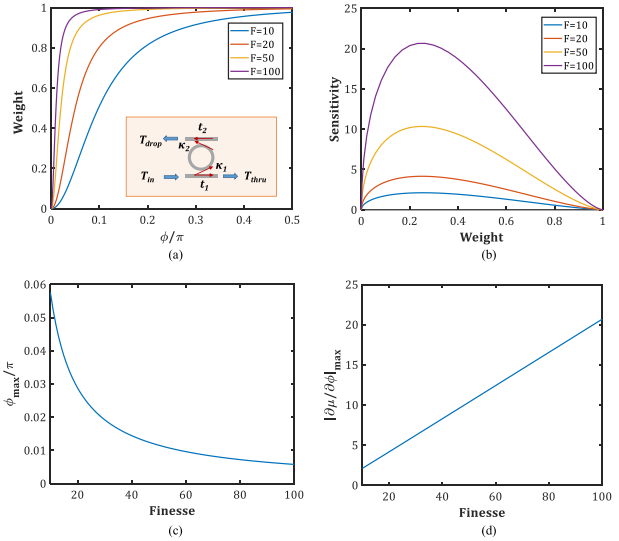

Fig. 9. ***(a) Weight definition of the through port of an add–drop MRR for critical coupling condition with various finesse. (b) Sensitivity of the weights for different finesse. (c) Optical phase shift at maximum sensitivity point. (d) Maximum sensitivity of weights as function of finesse. A linear trend is observed.***

on weight resolution due to thermal crosstalk. We focus on the thermo–optic phase shifting effect since it is a lossless tuning mechanism but imposes the most thermal impairments. We define the MRR weight sensitivity and discuss an approach to increase system stability. The ability to utilize only nonnegative values for training weight factors opens new opportunities to refine the ring locking scheme in the analog domain. A new class of highly accurate yet scalable OMMs that are based on add–drop MRRs for deep learning can thus be obtained.

1) Thermal-Crosstalk Restricted Weight Resolution:

a) Weight definition and sensitivity of add–drop MRR: An add–drop ring resonator refers to a circular ring structure that couples to two straight waveguides, as schematically shown by the inset in Fig. 9(a). The optical transfer function of the drop and through port can be expressed as [64]

$$D(\phi) = \left| \frac{-\kappa_1 \kappa_2 L^{0.25} \exp(-j\phi/2)}{1 - t_1 t_2 \sqrt{L} \exp(-j\phi)} \right|^2 \tag{5}$$

and

$$T(\phi) = \left| \frac{t_1 - t_2 \sqrt{L} \exp(-j\phi)}{1 - t_1 t_2 \sqrt{L} \exp(-j\phi)} \right|^2 \tag{6}$$

where t_1 and κ_1, t_2 and κ_2 are the self-coupling and cross-coupling coefficient for the input and drop coupling region, respectively. L is the round-trip optical power attenuation of the ring. We assume $t^2 + \kappa^2 = 1$ which allows the loss introduced by the couplers to be included in L. ϕ is the relative optical phase shift inside the ring

$$\phi = \frac{(\lambda - \lambda_{\text{res}})}{\text{FSR}} \times 2\pi \tag{7}$$

where λ_{res} is the ring resonance wavelength and FSR is the free spectral range of the resonance spectrum. We define the weighting function, μ, using the through port of the add–drop MRR, considering the negligible through loss and its flexibility in cascading. Equation (6) can be rewritten as

$$T(\phi) = \frac{T_0 + \left(\frac{2F}{\pi} \sin\left(\frac{\phi}{2}\right)\right)^2}{1 + \left(\frac{2F}{\pi} \sin\left(\frac{\phi}{2}\right)\right)^2} \tag{8}$$

where

$$T_0 = \frac{(t_1 - t_2\sqrt{L})^2}{(1 - t_1 t_2 \sqrt{L})^2} \tag{9}$$

and F is the finesse of the ring, given by

$$F = \pi \frac{\sqrt{t_1 t_2 \sqrt{L}}}{1 - t_1 t_2 \sqrt{L}} \approx \frac{\text{FSR}}{\Delta\lambda_{3\text{ dB}}} \tag{10}$$

where $\Delta\lambda_{3\text{ dB}}$ is the optical bandwidth of microring. Note that the approximation in (10) holds only when $F \gg 1$. Under critical coupling, the coupled power is equal to the power loss in the ring cavity, i.e., satisfying the relation $t_1 = t_2(L)^{1/2}$, hence the transmission drops to zero, $T_0 = 0$. Therefore, the design to operate at critical coupling mode enables the maximum extinction ratio (i.e., dynamic range) for the power transfer at the through port. The weighting function, μ, can thus be given as

$$\mu(\phi, F) = \frac{T(\phi)}{T(\pi)} = \frac{T_0 + \left(\frac{2F}{\pi} \sin\left(\frac{\phi}{2}\right)\right)^2}{T_0 + \left(\frac{2F}{\pi}\right)^2} \times \frac{1 + \left(\frac{2F}{\pi}\right)^2}{1 + \left(\frac{2F}{\pi} \sin\left(\frac{\phi}{2}\right)\right)^2}. \tag{11}$$

We define F as a variable of μ since finesse stands out as a key parameter of both the ring sensitivity and the scalability of number of neurons [77]. The finesse is a measure of the sharpness of resonances relative to their spacing (FSR) and represents, within a factor of 2π, the number of round trips made by light in the ring before its energy is reduced to $1/e$ of its initial value [63]. Therefore, from this point of view, the round-trip loss, L, as well as the coupling coefficients in the coupling regions of the ring, t_1 and t_2, are loss factors that can be manipulated to alter F, which is also reflected in (10). For datacom applications, MRRs are generally designed with radii in the region of 5–10 μm to avoid undesired high bending losses (i.e., radiation and scattering) while maintaining reasonably large FSR, therefore limiting the finesse to the order of tens [64]. Special designs, however, can lead to finesse with values of a few hundreds [90], [91]. Further details are discussed in Section IV-C1. In Fig. 9(a), we plot the weight factor as a function of ϕ for F values in the range of 10–100 (10, 20, 50, and 100), assuming the critical coupling operation ($T_0 = 0$).

It is not a surprise to see that a sharper resonance, i.e., larger F, gives rise to a more abrupt change in the weight

as a function of ϕ. It has been shown that the thermo-optic response (i.e., optical phase shift, $\Delta\phi$) of the microring is a linear function of heating power (ΔP) [82]. However, for a thermal perturbation (ΔP), $\Delta\mu$ varies depending on the weight μ, due to the nonlinear behavior of the optical transfer function. We thus define the sensitivity of the weights as the slope of the weight

$$\frac{\Delta\mu}{\Delta\phi}\approx\frac{\partial\mu}{\partial\phi}=\frac{(1+a^2)(1-T_0)}{T_0+a^2}\frac{0.5a^2\sin\phi}{\left(1+\left(a\sin\frac{\phi}{2}\right)^2\right)^2} \tag{12}$$

where $a=2F/\pi$. Combining with (11), we can plot the sensitivity as a function of weight for various values of F as in Fig. 9(b) for critical coupling ($T_0=0$). One can see that a lower sensitivity exists in the weighting function with a smaller F, where the change in weight is milder over $\Delta\phi$. This can be understood by realizing that a lower F results in a wider resonance linewidth, hence the weight has a smaller gradient as seen in Fig. 9(a). The optical phase settings at the maximum sensitivity (i.e., $\partial^2\mu/\partial\phi^2=0$) as a function of F is further given by

$$\phi_{\max}=2\tan^{-1}\sqrt{\frac{3a^2+2-\sqrt{9a^4+4a^2+4}}{a^2-2+\sqrt{9a^4+4a^2+4}}} \tag{13}$$

and illustrated in Fig. 9(c), indicating the weight variations are most sensitive close to the resonance, which agrees with the trend illustrated in Fig. 9(a) due to the nonlinear power transfer in MRRs. Fig. 9(d) indicates that the maximum sensitivity of the weight has a linear dependence on the finesse of the MRR, again showing that larger finesse leads to worse sensitivity. To facilitate the quantitative analysis on the bounded effective resolution, we use a first-order Taylor expansion of $\partial\mu/\partial\phi$ assuming that $a^2\gg 1$ (see Appendix II) to show this. The result is

$$\left|\frac{\partial\mu}{\partial\phi}\right|_{\max}\approx\frac{9}{16\sqrt{3}}a=\frac{3\sqrt{3}}{8\pi}F=0.2067\text{ F}. \tag{14}$$

b) Thermal crosstalk induced weight error: Thermal crosstalk occurs due to the proximity of rings to each other. The linear dependence of the temperature changes on the heater power results in a linear perturbation relation of the ring's temperature

$$\Delta T_i^{x\text{ talk}}=\sum_{j=1,j\neq i}^{N}\chi_j P_{H,j}(\mu_j) \tag{15}$$

where P_H is the heating power of other rings for setting their corresponding weights. This change of temperature translates into a change in the optical phase inside the ring

$$|\Delta\phi|=|\Delta\lambda_{\text{res}}|\frac{2\pi}{\text{FSR}}\approx 0.07\times|\Delta T^{x\text{talk}}|\times\frac{2\pi}{\text{FSR}}. \tag{16}$$

We use 0.07 nm/K as the typical resonance thermal sensitivity of silicon microrings [82]. Thermal crosstalk can be considered as a biased (deterministic) perturbation; hence, it affects the average value of the error, $|\overline{\Delta\mu}|$. Since the optical phase shift due to thermal crosstalk is a direct consequence of the weight of other rings whereas the weight sensitivity is dependent on the weight of interest, these two factors are uncorrelated and both can simultaneously occur at their worst cases. Therefore, the maximum weight error due to thermal crosstalk can be written as

$$\max|\overline{\Delta\mu}|=\left|\frac{\partial\mu}{\partial\phi}\right|_{\max}\times|\Delta\phi|_{\max}=0.091\times\frac{|\Delta T^{x\text{talk}}|_{\max}}{\Delta\lambda_{3\text{ dB}}}. \tag{17}$$

Considering adjacent MRR elements as thermal crosstalk sources and that the maximum phase shift inside each adjacent ring is π, the maximum temperature change due to thermal crosstalk from an adjacent ring can be given as

$$\Delta T_{\max}^{x\text{talk}}\approx 7.143\text{ FSR}\times\alpha_T \tag{18}$$

where α_T is the fraction of the thermal energy from adjacent rings. The weight error then aggregates as

$$\max|\overline{\Delta\mu}|=0.65\text{ F}\times\sum_i\alpha_{T,i}. \tag{19}$$

The solution of heat diffusion in 2-D space of the chip has a form of [82]

$$q(r)=q_{\text{ring}}\times\frac{\ln\left(\frac{R_\infty}{r}\right)}{\ln\left(\frac{R_\infty}{R}\right)} \tag{20}$$

where q is thermal energy density (proportional to the change in temperature at each location), r is the distance

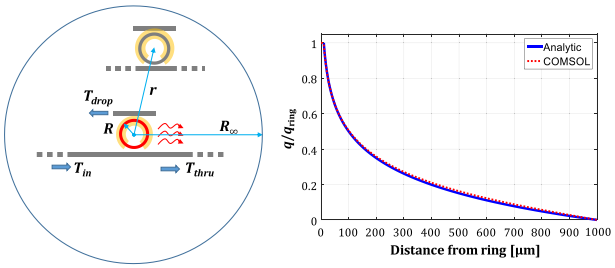

Fig. 10. ***(a) Schematic of thermal crosstalk between adjacent MRRs. R_∞ denotes the boundary of the chip. Thermal crosstalk arises from in-plane diffusion of heat and gets worse at closer proximity. (b) Comparison of analytic 2-D equation for heat diffusion with finite element results in COMSOL. The logarithmic behavior for the heat diffusion is confirmed. Note that the heat density, q, is proportional to the temperature change, ΔT, and can be considered a measure of thermal crosstalk.***

to the crosstalk source, R is the radius of the ring, and R_∞ can be viewed as the boundary of the chip, as shown in Fig. 10(a). Fig. 10(b) shows the validation of this analytic equation with COMSOL simulation [82], [92] for $R = 10$ μm and $R_\infty = 1$ mm. As expected, the heat density decreases at farther distances from the MRR's heater, but the 2-D heat diffusion shows a rather strong thermal crosstalk impact (e.g., 50% at 100-μm proximity). Note that in an actual photonic chip the heaters are most commonly located on top of the MRR so that the heat can also diffuse vertically. Since the thickness of the heater, t_H, is typically much smaller than the footprint of the heater (≈100 nm [93]), most of the heat generated by the heater diffuses vertically (out of plane) instead of horizontally (in-plane). Therefore, the fractional in-plane heat crosstalk from one ring to another can be estimated by

$$\alpha_T \approx \frac{t_H}{2R} \times \frac{\ln\left(\frac{R_\infty}{r}\right)}{\ln\left(\frac{R_\infty}{R}\right)} \tag{21}$$

and thus

$$\max|\overline{\Delta\mu}| = 0.65F \times \frac{t_H}{2R} \times \sum_i \frac{\ln\left(\frac{R_\infty}{r_i}\right)}{\ln\left(\frac{R_\infty}{R}\right)}. \tag{22}$$

c) Weight resolution: The resolution determines the minimum possible steps for setting weights with the highest certainty. If μ is the calibrated weight in the ideal case and $\hat{\mu}$ is the weight in the presence of perturbations, we can write

$$\hat{\mu} = \mu + \Delta\mu(t) = \mu + \overline{\Delta\mu} + \delta\mu(t) \tag{23}$$

where $\Delta\mu(t)$ is the error of the weight. This error can be decomposed into a stationary (deterministic) average denoted by $\overline{\Delta\mu}$ and a random noise like term denoted by $\delta\mu(t)$. We consider the resolution is set by the maximum root-mean-square error given by $\max|\Delta\mu(t)| = \max|\overline{\Delta\mu}| + \sigma_\mu/2$, where $\sigma_\mu^2 = \overline{\delta\mu^2(t)}$ is the standard deviation of the noise-like error. The resolution is then written as

$$\text{Resolution} = \frac{1}{\max|\Delta\mu(t)|} = \frac{1}{\max|\overline{\Delta\mu}| + \frac{\sigma_\mu}{2}}. \tag{24}$$

In such a system, it is reasonable to assume the thermal crosstalk induced error (i.e., $\overline{\Delta\mu}$) is dominant over the photo-diode noise, σ_μ. For an MRR element in an array, its two adjacent rings are considered as the dominant sources of thermal crosstalk. Therefore, referring to (22), we can plot the contours of effective bit resolution for an MRR unit as a function of both the unit pitch and its finesse for $R = 10$ μm and $R_\infty = 1$ mm, as shown in Fig. 11.

Note that this model is more accurate for small thermal perturbations; however, the combination of (22) and (24) still serves as a qualitative analysis on how the pitch size of MRR weighting elements and their finesse bound the effective resolution, even when the thermal crosstalk is strong. Feedforward calibration can somewhat alleviate the thermal crosstalk restrictions, yet the calibrated system accuracy heavily depends on the weight settings of adjacent MRR units. A scalable OMM with the capability of high resolution thus calls for a new design approach and the capabilities of computing using only nonnegative weight factors open up a new design philosophy, as discussed in the following section.

2) Hitless Weight-and-Aggregation Architecture: We propose a codesigned architecture for optical matrix multipliers which are specially customized for highly accurate, scalable, and nonnegative weight matrices. The hitless weight-and-aggregation design essentially describes an interconnect architecture that allows computational nodes (neurons) to carry arbitrary input vectors and to be independently weighted and summed. Such a many-to-one network is formed on the basis of channelization of the spectrum, creating physical and logical connections between input and output vectors. We put forward a hitless weighting structure by employing the colored channels in parallel rather than cascading them. This design isolates each weight on each connection and makes the tuning of MRR filters truly independent, i.e., not interfering with other channels. Such a hitless design also decouples the weighting and summation functions by allocating dedicated functional blocks, both of which employ MRR units, thus allowing independent optimization to decouple the constraint between the scalability of neurons and the weight sensitivity. The nonnegative weights are defined using the optical transfer function of the MRR through port, while the drop port is used as a monitoring outlet to provide real-time feedback for the weight control loop.

An $M \times N$ OMM consisting of $MN \times 1$ vector multipliers is illustrated in Fig. 12. Distinct continuous-wave (CW) wavelengths (representing N neurons) can be implemented by either M sets of N wavelength-multiplexed

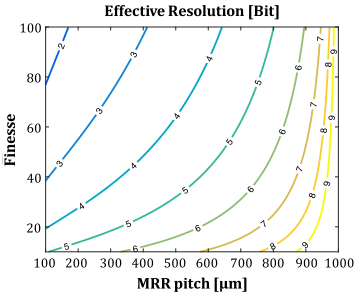

Fig. 11. ***Contours of effective bit resolution for the weight of MRR due to thermal crosstalk as a function of finesse and proximity.***

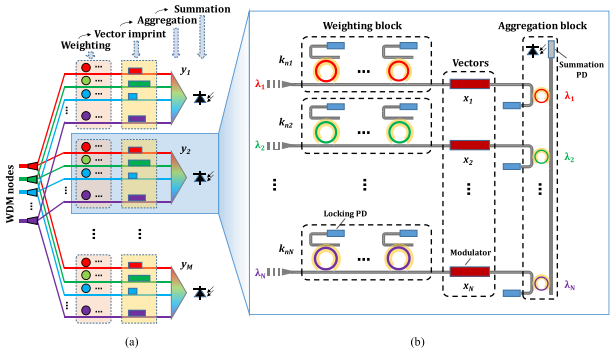

Fig. 12. ***(a) Hitless weight-and-aggregation architecture for M × N vector–matrix multiplier. (b) One unit out of M for N × 1 vector to be multiplied by 1 × N matrix.***

laser arrays [94] or optical frequency comb lasers [33], [95], or one set of lasers passively split into M copies. The nonnegative weight factors obtained from the trained matrix parameters are mapped to the coupling ratios and imprinted to the CW signals using multiring weighting blocks. The colored signals that carry the same set of weights are routed to all outputs. The N input vectors are then formed by a set of intensity modulators to the fanned-in WDM signals, before combining to form the M output vectors. The aggregation will be performed by another dedicated set of N high-finesse MRRs that is critically coupled to the WDM bus waveguide. The wavelength-multiplexed data streamed into the bus are optically summed by a PD, in which the photocurrent represents the total optical power. The M output vectors are then sent for nonlinear processing. Design considerations for each functional block are detailed in the following sections.

a) Hitless architecture for nonnegative weight factors: The design philosophy for the MRR-based weighting block is different from the conventional approach, in which tuning a filter in a link where WDM signals coexist controls the power coupling of the desired wavelength. The drop spectrum of such an MRR filter also sees other channels on the bus and thus the tuning inevitably interferes with adjacent channels. Such interference not only limits the weight tuning range but also acts as an unbiased perturbation to the weight that bounds the resolution for the nonnegative OMM system. Thus, a large channel spacing is required which trades off the system scalability.

Instead of utilizing the cascading layout of MRRs, the hitless design exploits a parallel arrangement of the weighting filters, shown in Fig. 12. This strategy stabilizes the weighting block within each wavelength branch before multiplexing onto the WDM bus, ensuring full tuning independence. Therefore, the design considerations for the MRR weighting filters can be narrowed down to a sole factor, i.e., sensitivity. As defined in (8) and Fig. 9(b), a small finesse is favored. Note that a trade-off exists since higher optical phase change is required to set the MRR to a specific weight for smaller finesse, which translates into higher heating power and, in turn, makes the thermal crosstalk worse. However, (9) still provides the worst possible scenario for the thermal crosstalk effects.

While the filter-through port is used to define the weighting function, the drop port connects to a monitor PD, shown in Fig. 13(a), providing a highly accurate feedback control loop for precise ring power locking. Fig. 13(b) plots the normalized monitor power for this structure as a function of ϕ, together with the corresponding weight factors. The locking accuracy could be compromised at power levels approaching zero (weight factors approaching one), given the existence of PD shot noise. To obtain a more linear transmission response, the ring spectrum tail can be omitted at the sacrifice of a slightly reduced weighting range.

The precise locking scheme would require a calibrated process, which sets up a lookup table (LUT) that maps the weight factor to the monitored optical power for each filter. By periodically polling the power monitor and comparing it to the LUT, the locking scheme can effectively offset thermal perturbations, including on-chip thermal crosstalk, and ambient temperature fluctuations. The locking accuracy, which could translate into weight resolution, can be limited by the PD shot noise, the finite precision that offers by the digital-to-analog convertor/analog-to-digital convertor (DAC/ADC), as well as the polling and locking rate.

b) Multiring weighting block for reduced sensitivity: By utilizing multiple MRR filters as illustrated in Fig. 13(c), the weight sensitivity can be further relaxed. The overall weighting function, μ_o, for n cascaded ring filters can be given as

$$\mu_o = \mu_1 \cdot \mu_2 \cdots \mu_n. \tag{25}$$

For simplicity, we assume $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, in which case μ is given by (11) and the phase settings are the same for all MRR filters. Fig. 14(a) plots μ_o as a function of ϕ, for $n = 1 - 5$, with $F = 10$. It can be seen that the weighting function gets increasingly linear as n increases.

We can analyze two cases for the weight sensitivity of the multiring system: 1) one ring is perturbed thermally and 2) all rings are perturbed thermally at the same time,

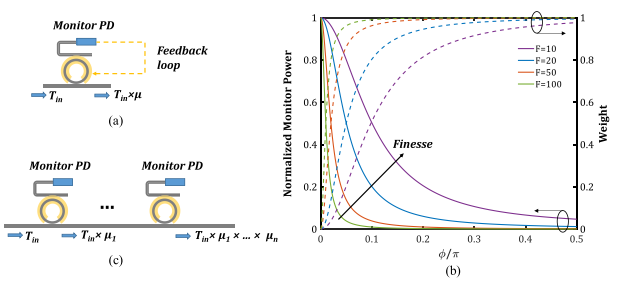

Fig. 13. ***(a) Single MRR weighting element with monitor PD. (b) Normalized monitor power as well as corresponded weight factor as a function of ϕ. (c) Multi-MRR weighting element.***

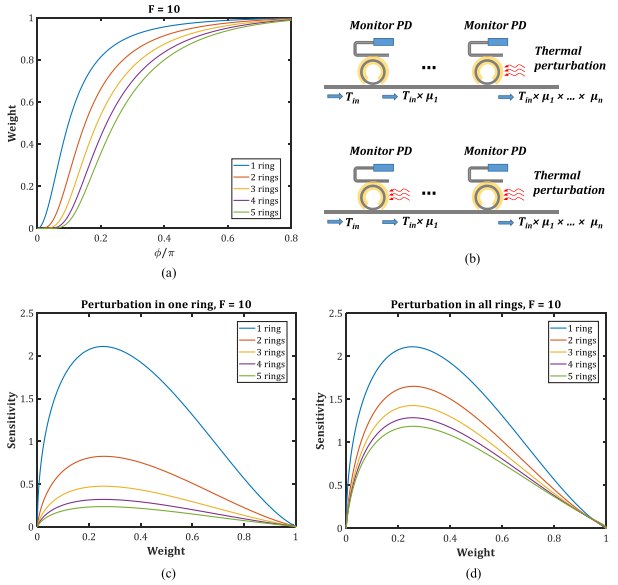

Fig. 14. *(a) Weight as a function of ϕ, for $n = 1 - 5$, with $F = 10$. (b) Illustration for one perturbed ring and all perturbed rings in a multiring weighting block. Weight sensitivity as a function of ϕ with thermal perturbation in (c) one ring and (d) all rings simultaneously.*

shown in Fig. 14(b). When the OMM setting leads to one or multiple heat sources on a chip, the dominant thermal effect is considered to be from adjacent rings. It is thus reasonable to take the one perturbed ring as the lower boundary for weight sensitivity. We have

$$\frac{\partial \mu_o}{\partial \phi} = \frac{\partial \mu_1}{\partial \phi} \cdot \mu_2 \cdots \mu_n. \tag{26}$$

This can be readily solved by referring to (12). The weight sensitivity with thermal perturbation in one ring can thus be plotted and is shown in Fig. 14(c), in which the single ring case is included for direct comparison. It can be seen that the two-ring system suppresses weight sensitivity significantly, but the trend continues with a decreasing decrement when the number of rings increases. For the case that thermal perturbation occurs in all rings, we have

$$\frac{\partial \mu_o}{\partial \phi} = \sum_{i=1}^{n} \frac{\mu_o}{\mu_i} \frac{\partial \mu_i}{\partial \phi} = \frac{\partial (\mu^n)}{\partial \phi}. \tag{27}$$

Fig. 14(d) plots this case representing the upper boundary for weight sensitivity. It can be seen that the system still gains tolerance to thermal perturbation compared to the single ring case. Considering the additional cost, footprint, and complexity introduced by the multiring system, the lower number two is preferred. Therefore, for the implementation of an $M \times N$ vector–matrix multiplier, the total number of MRRs is $3M \cdot N$ including both weighting MRRs and aggregation MRRs. The total number of PDs is M.

Although the multiring system exhibits lower weight sensitivity, overcoming the limitation of the finite precision for the DAC with which an optical phase can be set is still a challenge. In an n-ring weighting block, the minimum step in the weight, $\delta\mu$, bounded by the DAC resolution for setting the optical phase of each ring yields a weight $\hat{\mu} = \mu \pm \delta\mu$; hence the overall weight is $\hat{\mu}_o = (\mu \pm \delta\mu)^n \approx \mu^n(1 \pm n\delta\mu/\mu)$. Therefore, the error given by $n\mu^{n-1}\delta\mu$ can be at its worst (i.e., $n\delta\mu$) when μ is close to 1. A smaller error than $\delta\mu$ is achieved only for weights for which $n\mu^{n-1} < 1$. For a two-ring weight block, the worst error is $2\delta\mu$ which can occur for any weight.

c) Aggregation and summation: In contrast to the MZI-based OIU for matrix multiplication where the input vectors are imprinted before feeding into the OIU [27], we process the vector imprint after the weighting stage. This is because the weight factor, i.e., coupling ratio, is locked by the dropped power as illustrated in Fig. 13, and the streamed input vectors with power fluctuations would deteriorate the locking accuracy. Therefore, the proposed processing flow as shown in Fig. 12 resolves this issue. The input vectors are imprinted via high-speed intensity modulators [96]. A linear intensity modulator, such as the Mach–Zehnder modulator, is favored [52]. As we analyzed in the following section, high computation accuracy can be obtained when the input vectors have the same resolution as the weights.

The weighted input vectors can subsequently be aggregated into the WDM bus through dedicated ring filters. As shown in Fig. 15(a), the locking scheme for the aggregation MRRs operates differently, where the through power is always locked at the minimum state for a total power drop. This nontunable feature ensures the maximal spectral efficiency regarding the number of wavelengths that can reside in the WDM bus.

Since the aggregation ring filters act only as wavelength multiplexers, a large finesse is favored in order to achieve high scalability in the number of wavelength channels, i.e., the number of neurons. For a given finesse, the number of channels that can be carried within one FSR is determined by the channel spacing. A tradeoff exists for the channel spacing as it also determines the interchannel crosstalk when the dropped signals pass through

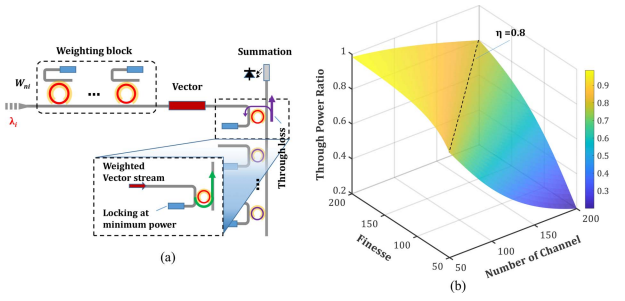

Fig. 15. *(a) Operating principle of the aggregation MRRs. (b) Through power ratio as a function of both finesse and number of channels.*

neighboring rings toward the summation PD on the bus. This leads to a through loss as illustrated in Fig. 15(a). We can rewrite (7) as

$$\phi = \frac{(\lambda - \lambda_{\text{res}})}{\text{FSR}} \times 2\pi = \frac{2\pi}{N_\lambda} \tag{28}$$

where $(\lambda - \lambda_{\text{res}})$ and N_λ are the channel spacing and number of channels, respectively. A large portion of the power loss gets dropped to the locking PD. This, however, does not compromise the weighting resolution. If we limit the through power ratio to η, we have

$$N_\lambda = 2\pi / \cos^{-1}\left(1 - \frac{2}{\frac{4F^2}{\pi^2}} \frac{T_0\left(1+\frac{4F^2}{\pi^2}\right) - \left(T_0 + \frac{4F^2}{\pi^2}\right)\eta}{\left(T_0+\frac{4F^2}{\pi^2}\right)\eta - \left(1+\frac{4F^2}{\pi^2}\right)}\right). \tag{29}$$

We can then plot a 2-D contour for η as a function of both finesse and number of channels, as shown in Fig. 15(b). Here, we assume the induced loss is dominated by the adjacent channel. It can be seen that for $\eta = 0.8$, which translates into ~1 dB through loss, $N_\lambda \approx F$. It should be noted that the insertion loss for all wavelength channels should be equalized by adjusting the individual input power, in order to allow each neuron to have the same maximum weight at summation. In addition, due to the multiring weighting block, the system can achieve higher order crosstalk suppression for the "0" weight.

B. System-Level Codesign

In order to take full advantage of both the optical speedup and electronic manipulation of the parallelism and memory, interactions between the two technologies require careful attention, especially when one processes digital signals and the other analog signals. We identify the system-level challenges for the codesign as follows.

1) Computation breakdown to match the interface. Processing a DNN may require matrix–vector multiplications for ultra-large matrices and vectors. The electronic circuitry should preprocess the DNN, breakdown the computation to smaller matrix–vector multiplication instances, send the request to a silicon photonic circuit, and post-process the results.
2) Minimization of the number of updates for the input matrix to the OMM. For each instance of matrix–vector multiplication requests, changing the values represented by the OMM microrings introduces a nonnegligible delay. Thus, to make the most of the high capacity of optical interconnects, it is desirable to have the elements of the input matrix to the OMM constant over a sequence of matrix–vector multiplication requests sent from the electronic device.
3) Analyzing the computation precision and nonnegative networks. As discussed in the previous sections, photonics is most suitable with nonnegative weights which can be directly mapped to the power ratios. The capability of defining weights using only nonnegative values would significantly simplify the design, fabrication, and control for the optical programmable processers. However, conventional training algorithms are developed using complementary (±) weight factors. Thus, it is important to investigate how the resolution level, nonnegative mode, and network size affect the accuracy of a neural network for a target task.
4) System-level scheduling and orchestration. To maximally utilize both types of devices, the latency of each device should be taken into account during the system-level scheduling and orchestration.

1) Fully Connected Layers: Computation Breakdown: For a fully connected layer h^{j+1} of size i_{j+1} from the previous layer h^j of size i_j, let $W^{j+1} \in \mathbb{R}^{i_{j+1} \times i_j}$ denote the weight matrix. Note that i_j and i_{j+1} can be much larger than N. Given an activation function Act() and bias b^{j+1}, the layer h^{j+1} can be computed as follows:

$$h^{j+1} = \text{Act}(g^{j+1} + b^{j+1}) \tag{30}$$
$$g^{j+1} = W^{j+1} \cdot h^j. \tag{31}$$

To compute g^{j+1} using the aforementioned PIC, we can partition the input into matrices of size $N \times N$ and vectors of size N as follows for $0 \le k \le i_{j+1}/N$:

$$\begin{pmatrix} g^{j+1}_{k+1} \\ g^{j+1}_{k+2} \\ \vdots \\ g^{j+1}_{k+N} \end{pmatrix} = \sum_{\ell=0}^{\frac{i_j}{N}} \begin{pmatrix} W^{j+1}_{k+1,\ell+1} & W^{j+1}_{k+1,\ell+2} & \cdots & W^{j+1}_{k+1,\ell+N} \\ W^{j+1}_{k+2,\ell+1} & W^{j+1}_{k+2,\ell+2} & & W^{j+1}_{k+1,\ell+N} \\ \vdots & & \ddots & \vdots \\ W^{j+1}_{k+N,\ell+1} & W^{j+1}_{k+N,\ell+2} & \cdots & W^{j+1}_{k+N,\ell+N} \end{pmatrix} \cdot \begin{pmatrix} h^j_{\ell+1} \\ h^j_{\ell+2} \\ \vdots \\ h^j_{\ell+N} \end{pmatrix}. \tag{32}$$

The overview of this approach is also depicted in Fig. 16.

The total number of multiplications required to compute layer h^{j+1} from h^j is $i_{j+1} \cdot i_j$. With the above approach using OMMs of width N, the total number of OMM requests is $\lceil i_{j+1}/N \rceil \cdot \lceil i_j/N \rceil$. This reduction by the factor of $1/(N^2)$ is achievable because there is no waste of operations associated with the partitioning.

2) Convolutional Layers: Minimization of the Reconfiguration of OMM Input Matrices: Fig. 17 shows the convolution part of convolutional layers computed using OMMs. The total number of multiplications required in computing one output channel is

$$i_{\text{in\_ch}} \cdot (W-2) \cdot (H-2) \cdot N^2 \tag{33}$$

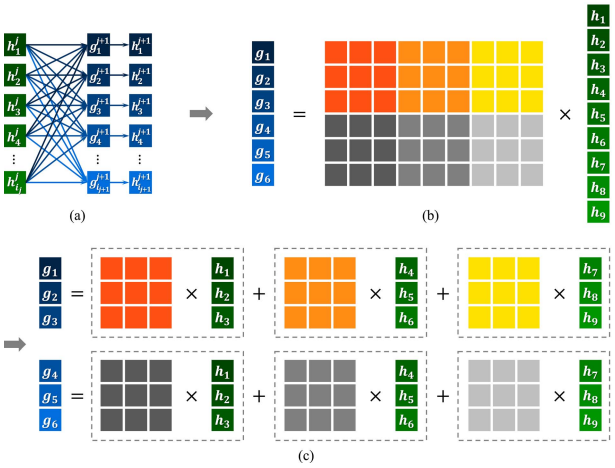

Fig. 16. *Computation for fully connected layers using OMMs. (a) Fully connected layers h^j (green) and h^{j+1} (blue). g^{j+1} is obtained as a result of the inner products between h^j and the weight vectors. h^{j+1} is obtained by applying the bias and activation to g^{j+1}. (b) Matrix–vector multiplication between the weight matrix (orange and gray) and h^j to obtain g^{j+1}. The superscripts are omitted for simplicity. (c) Computation equivalent to that of (b) but using OMMs with the input matrix size of* 3×3.

where $i_{\text{in\_ch}}$ denotes the number of input channels, W and H denote the width and height of an input channel, and N^2 represents the size of the convolution kernel. With the above approach, the total number of OMM requests for computing one output channel is $i_{\text{in\_ch}} \cdot (W-2) \cdot (H-2)$.

The above approach updates the matrix elements for each OMM request. On the other hand, we propose another approach illustrated in Fig. 18, which minimizes the number of updates of the input matrix for the OMM. This approach follows a similar direction to the weight

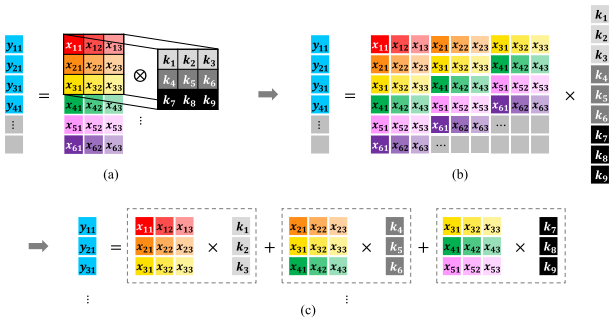

Fig. 17. *Computation for convolutional layers using OMMs. The first column of the output channel (nodes $y_{11}, \ldots$) and the first three columns of the input channel (nodes $x_{11}, \ldots$) are shown in the above illustration. (a) Convolutions on a single channel of the input layer. The convolution results over all channels in the input layer will be summed up and mapped to the output channel after the bias and activation are applied. (b) Conversion of the convolutions to a matrix–vector multiplication. The computation for one column in the output channel can be performed by a single matrix–vector multiplication for each input channel. (c) Computation equivalent to that of (b) but using OMMs with the input matrix size of* 3×3*, which equals the size of the convolution kernel.*

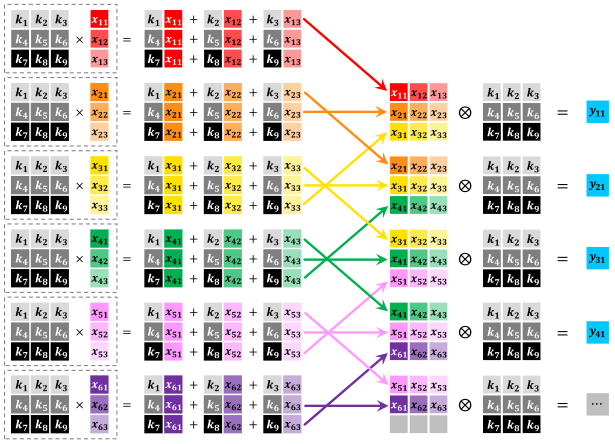

Fig. 18. *Proposed computation of convolutional layers using an OMM without updating its input matrix values.*

stationary optimization technique of ANN accelerators, where the weight values stay in the local register file of processing elements of the hardware accelerators [97]. The fundamental goal of this optimization is to minimize the time for processing elements to be reading the weights. In our codesigned system, the weights must be converted to analog signals and passed to the OMM to be set up for the computation. Thus, we aim at reducing the latency of the overall process by minimizing the number of OMMs input matrix updates. This can be achieved by mapping the convolution kernel itself to the OMM's input matrix, when the size of the OMM's input matrix is larger than or equal to that of the convolution kernel, which often ranges between 2×2 and 5×5. The convolution kernel weights form the input matrix and the network nodes form the input vectors. Then, the results of the first $N = 3$ matrix–vector multiplication instances in Fig. 10 contain the convolution result for y_{11}. The second, third, and fourth matrix–vector multiplication results contain the convolution result for y_{21}. Consecutive $N = 3$ results contain the convolution result for the corresponding output element. While processing the entire input channel, the input matrix for the OMM does not change. With this approach, the total number of OMM requests for one output channel is $i_{\text{in\_ch}} \cdot (W-2) \cdot H$.

3) Analysis of the Nonnegative Property and Resolutions: Most neural networks used in practice have both positive and negative input values, weights, and node values. Thus, feedforward propagation of these networks, either during the training or inference, requires matrix–vector multiplications with both positive and negative values. Then, it is of interest to consider a mapping between the values in the range of $[-1, 1]$ and $[0, 1]$ such that matrix–vector multiplication is preserved by this mapping. However, the theorem in Appendix I verifies that such mapping does not exist. There have been approaches to use only nonnegative input and weights to obtain a more understandable network with a slight decrease in the accuracy [98]. Another approach performs nonnegative matrix

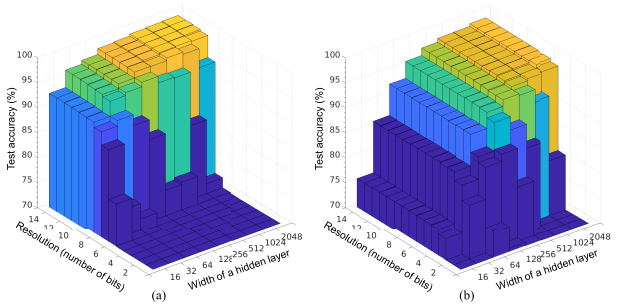

Fig. 19. *Test accuracy of MLPs for handwritten digit recognition with varying resolutions and network sizes. (a) Networks trained in the conventional mode using negative values, 0, and positive values in the computation. (b) Networks trained in the nonnegative mode, where only 0 and nonnegative values are used during matrix–vector multiplications.*

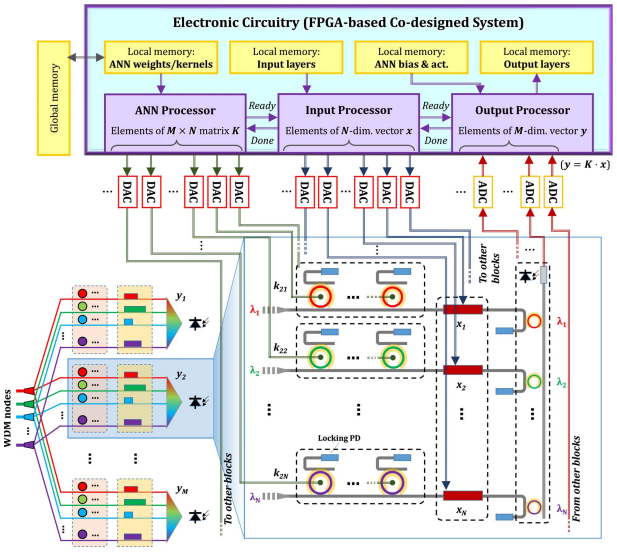

Fig. 20. *System-level overview for the proposed codesign approach. The FPGA-based electronic system (on the top) invokes and controls the optical system (in the bottom). The MRRs that receive electrical signals from DACs act as electrical-to-optical converters, whereas the summation PDs perform the optical-to-electrical conversions. The summed signals are connected to the FPGA via ADCs. Details regarding the memory systems, which depend on the specific application, are abstracted in this figure.*

factorization of the weights in order to reduce the input complexity, but the input values, in this case, can be both positive and negative [99].

To avoid matrix–vector multiplication with negative values, we train the neural networks using nonnegative input, weights, and nodes. In our experiment, we restrict not only the sign of the input and weights to be nonnegative but also the resolution used during inference. Fig. 19 shows the estimated inference accuracy of two-layer MLPs over a range of the resolution levels (the number of bits used to represent the input values and weights in a fixed-point format), and the network sizes (the number of nodes in the hidden layer of the MLP) trained in two different modes for the task of handwritten digit recognition: 1) conventional mode that supports negative input, weights, and nodes and 2) nonnegative mode that normalizes the input to [0, 1], and constrains the weights and nodes to be nonnegative. One network for each mode and each level of the network size was trained using the MNIST train data set [100], with 32-bit floating-point representation [101]. The input image contains 28×28 pixel values in the range of [0, 255], which were normalized to $[-1, 1]$ or [0, 1] depending on the training mode. For activation functions, tan$h$ was used in the hidden layer, and softmax was used in the output layer. After activation in the nonnegative mode, all negative values were rounded up to 0. All weights and biases were randomly initialized, and the weights for the nonnegative mode were initialized to [0, 1]. These weights and biases were updated using ADAM, which is a state-of-the-art stochastic back-propagation method [68].

Each of the trained networks was tested on the MNIST test data set, with both the input values and weights converted to the fixed-point representation for each resolution level. We note that one instance of a trained network with a given network structure does not represent the most optimized network of that structure. Nevertheless, all networks in this test case were trained using the same approach with similar optimization efforts, aside from the training time which increases for larger networks. Thus, we refer to these networks in order to practically and roughly estimate the performance trend over various network sizes, resolution levels, and the training mode. As shown in Fig. 19, the test accuracy has generally improved as the network width increased and as the resolution level was enhanced. It turns out that the accuracy of networks trained in the conventional mode was more affected by the restricted resolution, whereas the accuracy of those trained in the nonnegative mode was more affected by the network width.

The test accuracy achieved by nonnegative networks is lower than that by the conventionally trained counterpart, but a larger nonnegative network can sometimes outperform a smaller conventional network. During the training in the nonnegative mode, the biases and activation functions were allowed to take negative values because in this codesign approach only the matrix–vector multiplications will be offloaded from the electronic device to the optical device. This seems to have enabled the network to cut out less relevant, or negatively related connections and to focus on positively related ones, resulting in comparable accuracy for large nonnegative networks.

The issue of positive and negative inputs is an interesting example of the approach to optimization required for codesign. As mentioned in Section IV-A, photonics is implemented more readily with nonnegative values. This initial investigation indicates that, although in current practice both positive and negative values are used, using only nonnegative values for the matrix–vector multiplications can actually be advantageous in some circumstances.

4) System-Level Scheduling to Maximize the Throughput: To accelerate the inference process of a trained neural network with OMMs, an FPGA-based codesigned system breaks down the computation, sends matrix–vector multiplication requests to OMMs, and performs the remaining part of the computation including the nonlinear activation (which could also be done optically or via well-designed analog electronics as discussed in Section IV-C4). Fig. 20 illustrates the overview of the proposed codesigned system that contains three specialized processors: the ANN processor, the input processor, and the output processor. For each OMM request, the ANN processor sends the input $M \times N$ matrix K to the MRRs via DACs, and the input processor sends the input N-dimensional vector x to the modulators via DACs. The output processor collects the resulting M-dimensional vector y from the PDs via ADCs, and it also applies the bias and nonlinear activation function. The very recent demonstration on a 1-to-56-Gb/s ADC/DAC-based transceiver [102] paves the way for high-speed, low-energy ADC/DACs as the interface between the OMM and FPGA, without harming the throughput.

Although the computation complexity of an OMM is in O(1), the DAC, MRR configuration, ADC, and the computation on the FPGA consume nonnegligible latency. The goal of the system-level scheduling is to overlap these latencies to maximize the throughput. Fig. 21 shows abstract timing diagrams with pipelined executions by the ANN, input, and output processors. Fig. 21(a) illustrates the case of invoking a single OMM instance. As shown in Fig. 21(b), the latency T_L of a period between consecutive OMM invocations can be expressed as

$$T_L = T_M + T_{DA} + T_{AD} \tag{34}$$

where T_M denotes the latency of the DACs and MRR configuration, T_{DA} the latency of DACs and T_{AD} the latency of ADCs. This holds as long as the ANN processor's latency T_A does not exceed $T_{DA} + T_{AD}$, and similarly, the input and output processors' latencies T_I and T_O are less than or equal to $T_M + T_{AD}$ and $T_M + T_{DA}$, respectively.

When consecutive OMM instances contain the same input matrix elements so that it is not needed to reconfigure the MRRs, the latency T_L of the period can be expressed as

$$T_L = T_{DA} + T_{AD} \tag{35}$$

as shown in Fig. 21(c). In both cases of Fig. 21(b) and (c), the asymptotic throughput is proportional to $1/T_L$ and the number of OMM devices that can be processed in parallel, and is inversely proportional to the total numbers of OMM invocations for fully connected or convolutional layers which have been discussed in the previous sections.

C. Discussion

1) Silicon Ring Resonators (Finesse Versus Bandwidth): Silicon ring resonators with high finesse (up to a few hundreds) have been extensively demonstrated [90], [91]. However, these demonstrations aim for high-quality factors and tend to have a relatively small 3-dB bandwidth. For the aggregation ring filters in this OMM system, a large 3-dB bandwidth is an equally important factor that allows high data rate vectors to be fanned in, for high computational speeds. It would be preferable for the operation bandwidth of such an OMM to match that of the photo-detection rate (typically at 100 GHz).

The recent demonstration of a submicrometer-scale MRR shows great potential for the aggregation ring filters with high finesse and large bandwidth [103]. It features a 3-dB bandwidth of 100 GHz and a finesse of 116, supporting up to 116 wavelength channels given a 1-dB through loss budget as discussed in Section IV-A2c. This ultra-small ring resonator has the additional benefit of reducing the thermal tuning power, which is proportional to its size [103]. Another notable demonstration that combines an MRR-based filter with grating-assisted contra-directional couplers frees the constraint of FSR [104]. The addition of grating-assisted couplers provides an extra degree of freedom for longitudinal mode selectivity. This design, therefore, paves the way for independent optimization of the 3-dB bandwidth, and potentially enables full utilization of the transmission window in the silicon platform, yielding an extremely scalable OMM.

2) Optical Phase shifting Technology: Phase shifter technology is key in the OMM. Thermo-optic phase shifting is preferred since it is the most commonly applied lossless mechanism in the silicon platform. The induced

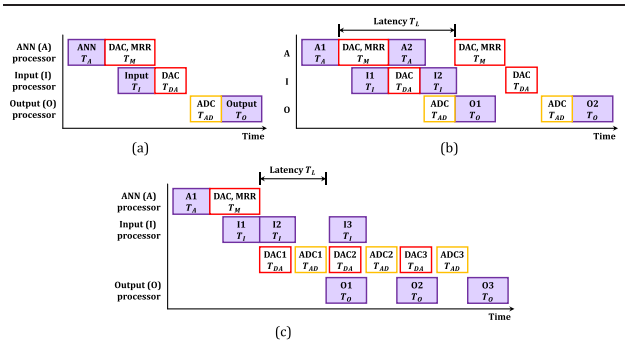

Fig. 21. ***(a) Timing diagram of invoking one OMM instance containing an input matrix and vector. (b) Timing diagram of invoking two OMM instances, each with an input matrix and vector. More invocations can be added on the right in a similar pattern. The latency of a period between consecutive invocations is denoted as T_L. (c) Timing diagram of invoking multiple OMM instances, where the first instance contains a new input matrix and vector and subsequent instances contain only new input vectors. The latency T_L has been reduced with respect to (b).***

on-chip thermal crosstalk can be reduced by adding isolation trenches [105]. In addition, a selective silicon etch can be applied to the silicon substrate to undercut the waveguides. The selective etch localizes the heat and improves heating efficiency [106]. The reduced heating power could, in turn, ameliorate on-chip thermal perturbations. The limited thermal frequency response (up to a few hundred kHz [82]) is, however, a limiting factor in latency, when dynamic reconfiguration for the OMM is required. For fast phase tuning, as aforementioned, electrooptic phase shifting leveraging the plasma dispersion effect is the most popular all-silicon technology [96]. It offers nanosecond-scale reconfiguration time, albeit with some performance penalty due to the electro-absorption loss. The E-O phase shifters would be straightforwardly included in the weighting blocks with additional considerations for the excess electro-absorption loss.

With the advances in heterogeneous integration technology, other materials can be introduced on the silicon platform. Notable examples include III–V materials [49], graphene [107], and nonvolatile phase-change materials (PCMs) [108]. III–V materials exhibit high electrooptic phase modulation efficiency, which can be effectively combined with silicon waveguides using wafer-bonding techniques [49]. Thin layers of graphene can be deposited on top of the Si waveguide [107], forming a capacitor that overlaps with the tail of the waveguide's optical mode. The application of voltage will then shift the Fermi level of graphene and enable inter-band transitions of charge carriers, and thus modulate the intensity of light traveling through the waveguide. The PCMs can introduce gigantic optical phase changes and most importantly, such phase changes are nonvolatile. This nonvolatility adds the capability of self-holding, maintaining optical states even in the absence of power input [109].

3) Power Consumption and Footprint of the OMM: The power consumption of the OMM is dominated by the tuning and locking of MRR elements. Current technology features a thermo-optic tuning efficiency of 1 nm/mW with doped-silicon micro-heaters [82], leading to small power consumption of a few mW per weighting MRR. Femtojoule-level depletion-mode modulators in vertically doped micro-disk structures [32], featuring low operating voltage (0.5 V_{PP}), offer the possibility for ultralow power electrooptic OMMs. The power consumption would then derive from the undesired leakage current, approximately $\sim$6 μW per element [32]. In future implementations, the phases could be set using the nonvolatile PCMs [109]. In that case, power would only be drawn during state transitions. A recent demonstration on a nonvolatile PCM-based photonic memory cell shows programming energy and time of only 680 pJ and 250 ns, respectively [110].

A number of wavelength locking schemes have been proposed, including the use of the photoconductive effect [87], small dithering signals [84], radio-frequency (RF) detection [86], additional partial drop rings [88], and monolithically integrated locking controllers in the 45-nm CMOS-SOI platform [85]. The locking power consumption has been demonstrated to be in the range of a few hundred μW [84], [85]. Furthermore, there has also been noteworthy research progress on a thermal MRRs that could significantly overcome the temperature sensitivity [111]–[113]. Here, the key idea is to introduce an upper cladding that has a negative thermo-optic coefficient to counteract the T-O effect of silicon. Titanium dioxide (TiO_2) holds the most promise as it exhibits a relatively strong negative thermo-optic coefficient and has been included in the CMOS-compatible fabrication process [112], [113]. This technique offers a path to extremely power-efficient OMM units.

Current implementation of MRR-based PICs for ON–OFF switching (two-state) applications normally features a pitch size of 100 μm [29]. Hundreds of MRR elements have been monolithically integrated on a single chip, within an area of a few tens of millimeter squares [29]. The temperature-insensitive MRRs could potentially reduce the footprint of the OMM significantly, even for high-resolution operations, as the pitch limitation due to thermal restrictions is offset. The size will then be merely limited by the pitch size of electrical bonding pads, which can be as small as 25–40 μm [114], thus enabling the footprint shrink of the OMM by over an order of magnitude.

4) Nonlinear Activation Function: To implement a full neural network, as aforementioned, a nonlinear activation function is required in addition to the linear OMM units. For a nonlinear activation function implemented in optics, there are generally two types, implemented using: 1) electrooptic nonlinearity and 2) all-optical nonlinearity. The former type requires first converting an optically weighted signal into the electrical domain and then triggering the nonlinear activation function to have an optical outcome. Examples include semiconductor excitable lasers [type C in Fig. 6(c)] [77] and electro-absorption modulators [115]. This type of solution might impair the processing speed and cascadability of neural networks due to the movement of charge carriers and the optical-to-electrical conversion noise. The latter, all-optical solution holds greater promise. The most commonly used optical nonlinearities are saturable absorption, such as in the use of monolayer graphene absorbers [116] and two-photon absorption [117], and bistability in nonlinear photonic crystals [118] and optical superlattices [119]. The nonlinearity of ring resonators can also be exploited [120]. Currently, the optical nonlinear activation function is an important research topic which could be used in order to enhance the throughput of an optical neural network, thus lowering the system latency and power consumption. However, the monolithic integration of these nonlinear units with OMMs, the efficiency of the nonlinear

modulation, and the operational speed and accuracy are open challenges [121].

While the development of an all-optical on-chip neural network represents a longer term goal, implementing the nonlinear activation function electrically is a promising alternative in the short term. The very recent work of building optical neural networks based on photoelectric multiplication also proposes to implement the nonlinear activation function in the electrical domain [122]. Very low power (femtojoule-scale) consumption is feasible with well-designed analog electronics.

V. CONCLUSION

Larger DNNs, in general, have higher expressiveness as a classification function. Theoretical analysis has also verified that both the depth and the width of neural networks contribute to their expressive power. It has been shown that complex functions expressed by DNNs cannot be approximated by any shallow neural network whose size is no more than an exponential bound [123], and also that certain classes of wide neural networks cannot be realized by any narrow network whose depth is no more than a polynomial bound [124]. These observations lead to the demand for the capability to efficiently process very deep or wide neural networks. The codesign approach addresses scalability (in terms of the size of neural networks) in two aspects: 1) the capability to decompose a large matrix–vector multiplication into smaller instances which significantly relaxes the requirement of photonic integrations and 2) a path to construct ultra-large scale OMMs using MRRs in the wavelength domain. This reduces the system decomposition complexity and, in turn, enables the handling of sophisticated concepts for future applications. In addition, the approach to manage the computation precision with nonnegative values can be utilized in any photonic systems, in order to reduce the implementation complexity and thus cost. This also facilitates the operation of different facets of validity in practical terms for OMMs as hardware accelerators in deep-learning applications.

In summary, efficient scaling of deep learning will require dedicated hardware accelerators. We have presented an overview of silicon photonics applications for deep learning and have analyzed opportunities for scalable codesigned multiwavelength microring silicon photonic architectures. ■

APPENDIX A

Theorem 1: Let $\Phi, \Omega \subset \mathbb{R}$ such that $\{-1, 0, 1\} \in \Phi$ and $\Omega \subset [0, +\infty]$. Then there exists no function $f : \Phi \to \Omega$ satisfying the following:

$$\text{For any } p_1, p_2 \in \Phi,\ f(p_1) + f(p_2) = f(p_1 + p_2) \tag{A1}$$

$$\text{For any } p_1, p_2 \in \Phi,\ f(p_1) \cdot f(p_2) = f(p_1 \cdot p_2). \tag{A2}$$

This also holds if $\Omega \subset (-\infty, 0)$.

Proof: If such a function f exists, it must satisfy the following:

$$f(1) + f(0) = f(1) \tag{A3}$$

$$f(1) \cdot f(-1) = f(-1) \tag{A4}$$

$$f(1) + f(-1) = f(0). \tag{A5}$$

Equation (A3) implies that $f(0) = 0$, and (A4) implies that $f(1) = 1$. Then, (A5) can be rewritten as

$$1 + f(-1) = 0. \tag{A6}$$

Thus, $f(-1) = -1$ but this value is not in the range Ω of function f. Therefore, such an f does not exist.

APPENDIX B

As discussed in Section IV-A1a, the maximum sensitivity of the weight in (11) occurs when

$$\tan \frac{\phi_{\max}}{2} = \sqrt{\frac{r(a)}{s(a)}} \tag{A7}$$

where

$$r(a) = 3a^2 + 2 - \sqrt{9a^4 + 4a^2 + 4} \tag{A8}$$

and

$$s(a) = a^2 - 2 + \sqrt{9a^4 + 4a^2 + 4}. \tag{A9}$$

Therefore

$$\sin \phi_{\max} = \frac{2\sqrt{r(a)s(a)}}{r(a) + s(a)} = \frac{1}{2a^2}\sqrt{r(a)s(a)} \tag{A10}$$

$$\cos \phi_{\max} = \frac{s(a) - r(a)}{s(a) + r(a)} = 1 - \frac{r(a)}{2a^2}. \tag{A11}$$

Plugging these back into the sensitivity function of (12) and assuming $T_0 \approx 0$ immediately yields

$$\left|\frac{\partial \mu}{\partial \phi}\right|_{\max} = \frac{1 + a^2}{a^2} \frac{4\sqrt{r(a)s(a)}}{(r(a) + 4)^2} \tag{A12}$$

Assuming that $a^2 \gg 1$, we see that

$$\frac{1 + a^2}{a^2} \approx 1 \tag{A13}$$

$$r(a) \approx 3a^2 + 2 - 3a^2\left(1 + \frac{2}{9a^2}\right) = \frac{4}{3} \tag{A14}$$

$$s(a) \approx a^2 - 2 + 3a^2\left(1 + \frac{2}{9a^2}\right) \approx 4a^2. \tag{A15}$$

Therefore

$$\left|\frac{\partial \mu}{\partial \phi}\right|_{\max} \approx \frac{9}{16\sqrt{3}}a. \tag{A16}$$

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[2] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.

[3] M. Bojarski *et al.*, "End to end learning for self-driving cars," 2016, *arXiv:1604.07316*. [Online]. Available: https://arxiv.org/abs/1604.07316

[4] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[5] W. Haensch, T. Gokmen, and R. Puri, "The next generation of deep learning hardware: Analog computing," *Proc. IEEE*, vol. 107, no. 1, pp. 108–122, Oct. 2019.

[6] M. Hu *et al.*, "Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," in *Proc. 53rd Annu. Design Autom. Conf. (DAC)*, Austin, TX, USA, 2016, pp. 1–6.

[7] R. A. Athale and W. C. Collins, "Optical matrix–matrix multiplier based on outer product decomposition," *Appl. Opt.*, vol. 21, no. 12, pp. 2089–2090, Jun. 1982.

[8] N. H. Farhat, D. Psaltis, A. Prata, and E. Paek, "Optical implementation of the Hopfield model," *Appl. Opt.*, vol. 24, no. 10, pp. 1469–1475, May 1985.

[9] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[11] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.

[12] D. E. Rumelhart and G. E. R. J. Hinton Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing—Explorations in the Microstructure of Cognition*, vol. 1, E. R. David, L. M. James, and C. P. R. Group, Eds. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.

[13] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals Syst.*, vol. 2, no. 4, pp. 303–314, 1989.

[14] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[15] Y. Lecun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[16] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[17] H. Bagherian, S. Skirlo, Y. Shen, H. Meng, V. Ceperic, and M. Soljacic, "On-chip optical convolutional neural networks," 2018, *arXiv:1808.03303*. [Online]. Available: https://arxiv.org/abs/1808.03303

[18] J. Bergstra *et al.*, "Theano: Deep learning on gpus with python," *J. Mach. Learn. Res.* vol. 3, no. 1, pp. 1–48, 2011.

[19] S. Chetlur *et al.*, "cuDNN: Efficient primitives for deep learning," 2014, *arXiv:1410.0759*. [Online]. Available: https://arxiv.org/abs/1410.0759

[20] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. S. Chung, "Accelerating deep convolutional neural networks using specialized hardware," *Microsoft Res. Whitepaper*, vol. 2, no. 11, pp. 1–4, 2015.

[21] S. Amiri, M. Hosseinabady, S. McIntosh-Smith, and J. Nunez-Yanez, "Multi-precision convolutional neural networks on heterogeneous hardware," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 419–424.

[22] S. K. Esser *et al.*, "Convolutional networks for fast, energy-efficient neuromorphic computing," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 41, pp. 11441–11446, Oct. 2016.

[23] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, no. 1, pp. 239–255, 2010.

[24] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[25] D. Casasent and A. Ghosh, "Optical linear algebra processors: Noise and error-source modeling," *Opt. Lett.*, vol. 10, no. 6, pp. 252–254, Jun. 1985.

[26] Y.-Z. Liang and H. K. Liu, "Optical matrix–matrix multiplication method demonstrated by the use of a multifocus hololens," *Opt. Lett.*, vol. 9, no. 8, pp. 322–324, Aug. 1984.

[27] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 7, pp. 441–446, Jul. 2017.

[28] Q. Cheng, M. Bahadori, Y.-H. Hung, Y. Huang, N. Abrams, and K. Bergman, "Scalable microring-based silicon Clos switch fabric with switch-and-select stages," *IEEE J. Sel. Topics Quantum Electron.*, vol. 25, no. 5, pp. 1–11, Sep. 2019.

[29] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent advances in optical technologies for data centers: A review," *Optica*, vol. 5, no. 11, pp. 1354–1370, Nov. 2018.

[30] A. E.-J. Lim *et al.*, "Review of silicon photonics foundry efforts," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 4, pp. 405–416, Jul. 2014.

[31] A. Rahim, T. Spuesens, R. Baets, and W. Bogaerts, "Open-access silicon photonics: Current status and emerging initiatives," *Proc. IEEE*, vol. 106, no. 12, pp. 2313–2330, Dec. 2018.

[32] E. Timurdogan, C. M. Sorace-Agaskar, J. Sun, E. S. Hosseini, A. Biberman, and M. R. Watts, "An ultralow power athermal silicon modulator," *Nature Commun.*, vol. 5, Jun. 2014, Art. no. 4008.

[33] S. Liu *et al.*, "High-channel-count 20 GHz passively mode-locked quantum dot laser directly grown on Si with 4.1 Tbit/s transmission capacity," *Optica*, vol. 6, no. 2, pp. 128–134, 2019.

[34] P. Chen, S. Chen, X. Guan, Y. Shi, and D. Dai, "High-order microring resonators with bent couplers for a box-like filter response," *Opt. Lett.*, vol. 39, no. 21, pp. 6304–6307, 2014.

[35] L. Chrostowski and M. Hochberg, *Silicon Photonics Design: From Devices to Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[36] J. Sun, E. Timurdogan, A. Yaacobi, E. S. Hosseini, and M. R. Watts, "Large-scale nanophotonic phased array," *Nature*, vol. 493, no. 7431, pp. 195–199, Jan. 2013.

[37] X. Qiang *et al.*, "Large-scale silicon quantum photonics implementing arbitrary two-qubit processing," *Nature Photon.*, vol. 12, no. 9, pp. 534–539, Sep. 2018.

[38] M. Yu *et al.*, "Silicon-chip-based mid-infrared dual-comb spectroscopy," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 1869.

[39] D. Pérez *et al.*, "Multipurpose silicon photonics signal processor core," *Nature Commun.*, vol. 8, no. 1, 2017, Art. no. 636.

[40] A. N. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 7430.

[41] H. Hu *et al.*, "Single-source chip-based frequency comb enabling extreme parallel data transmission," *Nature Photon.*, vol. 12, no. 8, pp. 469–473, 2018.

[42] Q. Cheng, S. Rumley, M. Bahadori, and K. Bergman, "Photonic switching in high performance datacenters [Invited]," *Opt. Express*, vol. 26, no. 12, pp. 16022–16043, Jun. 2018.

[43] L.-W. Luo *et al.*, "WDM-compatible mode-division multiplexing on a silicon chip," *Nature Commun.*, vol. 5, no. 1, 2014, Art. no. 3069.

[44] K. Kwon *et al.*, "128×128 silicon photonic MEMS switch with scalable row/column addressing," in *Conf. Lasers Electro-Opt., OSA Tech. Dig. (Online) (Opt. Soc. Amer.*, 2018, Paper SF1A.4

[45] J. Michel, J. Liu, and L. C. Kimerling, "High-performance Ge-on-Si photodetectors," *Nature Photon.*, vol. 4, Aug. 2010, Art. no. 527.

[46] D. Feng *et al.*, "High-speed GeSi electroabsorption modulator on the SOI waveguide platform," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 6, pp. 64–73, Nov. 2013.

[47] W. D. Sacher *et al.*, "Monolithically integrated multilayer silicon nitride-on-silicon waveguide platforms for 3-D photonic circuits and devices," *Proc. IEEE*, vol. 106, no. 12, pp. 2232–2245, Dec. 2018.

[48] O. Marshall, M. Hsu, Z. Wang, B. Kunert, C. Koos, and D. Van Thourhout, "Heterogeneous integration on silicon photonics," *Proc. IEEE*, vol. 106, no. 12, pp. 2258–2269, Dec. 2018.

[49] T. Komljenovic, D. Huang, P. Pintus, M. A. Tran, M. L. Davenport, and J. E. Bowers, "Photonic integrated circuits using heterogeneous integration on silicon," *Proc. IEEE*, vol. 106, no. 12, pp. 2246–2257, Dec. 2018.

[50] S. Chen *et al.*, "Electrically pumped continuous-wave III–V quantum dot lasers on silicon," *Nature Photon.*, vol. 10, no. 5, pp. 307–311, May 2016.

[51] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, "Micrometre-scale silicon electro-optic modulator," *Nature*, vol. 435, no. 7040, pp. 325–327, May 2005.

[52] C. Zhang, P. A. Morton, J. B. Khurgin, J. D. Peters, and J. E. Bowers, "Highly linear heterogeneous-integrated Mach–Zehnder interferometer modulators on Si," *Opt. Express*, vol. 24, no. 17, pp. 19040–19047, Aug. 2016.

[53] A. P. Ovvyan, N. Gruhler, S. Ferrari, and W. H. P. Pernice, "Cascaded Mach–Zehnder interferometer tunable filters," *J. Opt.*, vol. 18, no. 6, 2016, Art. no. 064011.

[54] M. Bahadori, S. Rumley, D. Nikolova, and K. Bergman, "Comprehensive design space exploration of silicon photonic interconnects," *J. Lightw. Technol.*, vol. 34, no. 12, pp. 2975–2987, Jun. 15, 2016.

[55] L.-W. Luo *et al.*, "High bandwidth on-chip silicon photonic interleaver," *Opt. Express*, vol. 18, no. 22, pp. 23079–23087, Oct. 2010.

[56] Q. Cheng *et al.*, "Ultralow-crosstalk, strictly non-blocking microring-based optical switch," *Photon. Res.*, vol. 7, no. 2, pp. 155–161, Feb. 2019.

[57] P. Dasmahapatra, R. Stabile, A. Rohit, and K. A. Williams, "Optical crosspoint matrix using broadband resonant switches," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 4, pp. 1–10, Jul. 2014.

[58] Q. Cheng, A. Wonfor, J. L. Wei, R. V. Penty, and I. H. White, "Demonstration of the feasibility of large-port-count optical switching using a hybrid Mach–Zehnder interferometer–semiconductor optical amplifier switch module in a recirculating loop," *Opt. Lett.*, vol. 39, no. 18, p. 5244, Sep. 2014.

[59] C. Mesaritakis, V. Papataxiarhis, and D. Syvridis, "Micro ring resonators as building blocks for an all-optical high-speed reservoir-computing bit-pattern-recognition system," *J. Opt. Soc. Amer. B, Opt. Phys.*, vol. 30, no. 11, pp. 3048–3055, Nov. 2013.

[60] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: An integrated network for scalable photonic spike processing," *J. Lightw. Technol.*, vol. 32, no. 21, pp. 4029–4041, Nov. 1, 2014.

[61] V. Ramaswamy and R. D. Standley, "A phased, optical, coupler-pair switch," *Bell Syst. Tech. J.*, vol. 55, no. 6, pp. 767–775, 1976.

[62] Q. Cheng, A. Wonfor, R. V. Penty, and I. H. White, "Scalable, low-energy hybrid photonic space switch," *J. Lightw. Technol.*, vol. 31, no. 18, pp. 3077–3084, Sep. 15, 2013.

[63] W. Bogaerts *et al.*, "Silicon microring resonators," *Laser Photon. Rev.*, vol. 6, no. 1, pp. 47–73, Jan. 2012.

[64] M. Bahadori *et al.*, "Design space exploration of microring resonators in silicon photonic interconnects: Impact of the ring curvature," *J. Lightw. Technol.*, vol. 36, no. 13, pp. 2767–2782, Jul. 1, 2018.

[65] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Phys. Rev. Lett.*, vol. 73, no. 1, pp. 58–61, Jul. 2002.

[66] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walsmley, "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, pp. 1460–1465, Dec. 2016.

[67] J. Carolan *et al.*, "Universal linear optics," *Science*, vol. 349, no. 6249, pp. 711–716, Aug. 2015.

[68] D. A. B. Miller, "Self-configuring universal linear optical component [Invited]," *Photon. Res.*, vol. 1, no. 1, p. 1, Jun. 2013.

[69] D. A. B. Miller, "Meshing optics with applications," *Nature Photon.*, vol. 11, no. 7, pp. 403–404, 2017.

[70] N. C. Harris *et al.*, "Quantum transport simulations in a programmable nanophotonic processor," *Nature Photon.*, vol. 11, no. 7, 2017, Art. no. 447.

[71] A. Ribeiro, A. Ruocco, L. Vanacker, and W. Bogaerts, "Demonstration of a 4 × 4-port universal linear circuit," *Optica*, vol. 3, no. 12, pp. 1348–1357, Dec. 2016.

[72] D. A. B. Miller, "Setting up meshes of interferometers–reversed local light interference method," *Opt. Express*, vol. 25, no. 23, pp. 29233–29248, 2017.

[73] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[74] T. Lindblad and J. M. J. K. Taylor, *Image Processing Using Pulse-Coupled Neural Networks*. New York, NY, USA: Springer, 1998.

[75] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Netw.*, vol. 10, no. 9, pp. 1659–1671, Dec. 1997.

[76] S. Thorpe, A. Delorme, and R. Van Rullen, "Spike-based strategies for rapid processing," *Neural Netw.*, vol. 14, nos. 6–7, pp. 715–725, Jul. 2001.

[77] F. De Lima, T. Ferreira, B. J. Shastri, A. N. Tait, M. A. Nahmias, and P. R. Prucnal, "Progress in neuromorphic photonics," *Nanophotonics*, vol. 6, no. 3, pp. 577–599, 2017.

[78] A. N. Tait *et al.*, "Microring weight banks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 22, no. 6, pp. 312–325, Nov. 2016.

[79] A. N. Tait, T. Ferreira De Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Continuous calibration of microring weights for analog optical networks," *IEEE Photon. Technol. Lett.*, vol. 28, no. 8, pp. 887–890, Apr. 15, 2016.

[80] A. N. Tait, T. F. De Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Multi-channel control for microring weight banks," *Opt. Express*, vol. 24, no. 8, pp. 8895–8906, Apr. 2016.

[81] A. N. Tait *et al.*, "Feedback control for microring weight banks," *Opt. Express*, vol. 26, no. 20, pp. 26422–26443, Oct. 2018.

[82] M. Bahadori *et al.*, "Thermal rectification of integrated microheaters for microring resonators in silicon photonics platform," *J. Lightw. Technol.*, vol. 36, no. 3, pp. 773–788, Feb. 1, 2018.

[83] A. F. M. Agarap, "On breast cancer detection: An application of machine learning algorithms on the wisconsin diagnostic dataset," in *Proc. 2nd Int. Conf. Mach. Learn. Soft Comput.*, Phu Quoc, Vietnam, 2018, pp. 5–9.

[84] K. Padmaraju, D. F. Logan, T. Shiraishi, J. J. Ackert, A. P. Knights, and K. Bergman, "Wavelength locking and thermally stabilizing microring resonators using dithering signals," *J. Lightw. Technol.*, vol. 32, no. 3, pp. 505–512, Feb. 1, 2014.

[85] C. Sun *et al.*, "A 45 nm CMOS-SOI monolithic photonics platform with bit-statistics-based resonant microring thermal tuning," *IEEE J. Solid-State Circuits*, vol. 51, no. 4, pp. 893–907, Apr. 2016.

[86] P. Dong *et al.*, "Simultaneous wavelength locking of microring modulator array with a single monitoring signal," *Opt. Express*, vol. 25, no. 14, pp. 16040–16046, Jul. 2017.

[87] H. Jayatilleka, H. Shoman, L. Chrostowski, and S. Shekhar, "Photoconductive heaters enable control of large-scale silicon photonic ring resonator circuits," *Optica*, vol. 6, no. 1, pp. 84–91, Jan. 2019.

[88] A. S. P. Khope *et al.*, "On-chip wavelength locking for photonic switches," *Opt. Lett.*, vol. 42, no. 23, pp. 4934–4937, 2017.

[89] W. A. Zortman, D. C. Trotter, and M. R. Watts, "Silicon photonics manufacturing," *Opt. Express*, vol. 18, no. 23, pp. 23598–23607, 2010.

[90] Q. Xu, D. Fattal, and R. G. Beausoleil, "Silicon microring resonators with 1.5-μm radius," *Opt. Express*, vol. 16, no. 6, pp. 4309–4315, 2008.

[91] A. Biberman, M. J. Shaw, E. Timurdogan, J. B. Wright, and M. R. Watts, "Ultralow-loss silicon ring resonators," *Opt. Lett.*, vol. 37, no. 20, pp. 4236–4238, Oct. 2012.

[92] *COMSOL 5.1*. Accessed: Mar. 19, 2019. [Online]. Available: https://www.comsol.com

[93] A. H. Atabaki, E. S. Hosseini, A. A. Eftekhar, S. Yegnanarayanan, and A. Adibi, "Optimization of metallic microheaters for high-speed reconfigurable silicon photonics," *Opt. Express*, vol. 18, no. 17, pp. 18312–18323, Aug. 2010.

[94] H. Nishi *et al.*, "Integration of eight-channel directly modulated membrane-laser array and SiN AWG multiplexer on Si," *J. Lightw. Technol.*, vol. 37, no. 2, pp. 266–273, Jan. 15, 2019.

[95] B. Stern, X. Ji, Y. Okawachi, A. L. Gaeta, and M. Lipson, "Battery-operated integrated frequency comb generator," *Nature*, vol. 562, no. 7727, pp. 401–405, Oct. 2018.

[96] J. Witzens, "High-speed silicon photonics modulators," *Proc. IEEE*, vol. 106, no. 12, pp. 2158–2182, Dec. 2018.

[97] Y.-H. Chen, J. Emer, and V. Sze, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," *SIGARCH Comput. Archit. News*, vol. 44, no. 3, pp. 367–379, 2016.

[98] J. Chorowski and J. M. Zurada, "Learning understandable neural networks with nonnegative weight constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 62–69, Apr. 2015.

[99] J. Flenner and B. Hunter. (2017). *A Deep Non-Negative Matrix Factorization Neural Network*. [Online]. Available: https://www.semanticscholar.org/paper/A-Deep-Non-Negative-Matrix-Factorization-Neural-Flenner-Hunter/ee8f9e9cf6ab21bb820589dc00803a39c5e30c63

[100] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[101] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[102] M. Pisati *et al.*, "A sub-250 mW 1-to-56Gb/s continuous-range PAM-4 42.5 dB IL ADC/DAC-based transceiver in 7 nm FinFET," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 116–118.

[103] D. Liu, C. Zhang, D. Liang, and D. Dai, "Submicron-resonator-based add-drop optical filter with an ultra-large free spectral range," *Opt. Express*, vol. 27, no. 2, pp. 416–422, Jan. 2019.

[104] N. Eid, R. Boeck, H. Jayatilleka, L. Chrostowski, W. Shi, and N. A. F. Jaeger, "FSR-free silicon-on-insulator microring resonator based filter with bent contra-directional couplers," *Opt. Express*, vol. 24, no. 25, pp. 29009–29021, Dec. 2016.

[105] X. Wu *et al.*, "Low power consumption voa array with air trenches and curved waveguide," *IEEE Photon. J.*, vol. 10, no. 2, Apr. 2018, Art. no. 7201308.

[106] A. Masood *et al.*, "Comparison of heater architectures for thermal control of silicon photonic circuits," in *Proc. 10th Int. Conf. Group IV Photon.*, 2013, pp. 83–84.

[107] C. T. Phare, Y.-H. Daniel Lee, J. Cardenas, and M. Lipson, "Graphene electro-optic modulator with 30 GHz bandwidth," *Nature Photon.*, vol. 9, no. 8, pp. 511–514, Aug. 2015.

[108] C. D. Wright *et al.*, "Integrated phase-change photonics: A strategy for merging communication and computing," in *Opt. Fiber Commun. Conf. (OFC), OSA Tech. Dig. (Opt. Soc. Amer.)*, 2019, Paper M1D.3.

[109] Q. Wang *et al.*, "Optically reconfigurable metasurfaces and photonic devices based on phase change materials," *Nature Photon.*, vol. 10, no. 1, pp. 60–65, Jan. 2016.

[110] X. Li *et al.*, "Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell," *Optica*, vol. 6, no. 1, pp. 1–6, Jan. 2019.

[111] J. Teng *et al.*, "Athermal Silicon-on-insulator ring resonators by overlaying a polymer cladding on narrowed waveguides," *Opt. Express*, vol. 17, no. 17, pp. 14627–14633, Aug. 2009.

[112] S. Feng *et al.*, "Athermal silicon ring resonators clad with titanium dioxide for 13 μm wavelength operation," *Opt. Express*, vol. 23, no. 20, pp. 25653–25660, Oct. 2015.

[113] B. Guha, J. Cardenas, and M. Lipson, "Athermal silicon microring resonators with titanium oxide cladding," *Opt. Express*, vol. 21, no. 22, pp. 26557–26563, 2013.

[114] J. H. Lau, "Recent advances and new trends in flip chip technology," *J. Electron. Packag.*, vol. 138, no. 3, 2016, Art. no. 030802.

[115] J. K. George *et al.*, "Neuromorphic photonics with electro-absorption modulators," *Opt. Express*, vol. 27, no. 4, pp. 5181–5191, Feb. 2019.

[116] Q. Bao *et al.*, "Monolayer graphene as a saturable absorber in a mode-locked laser," *Nano Res.*, vol. 4, no. 3, pp. 297–307, 2011.

[117] R. W. Schirmer and A. L. Gaeta, "Nonlinear mirror based on two-photon absorption," *J. Opt. Soc. Amer. B, Opt. Phys.*, vol. 14, no. 11, pp. 2865–2868, Nov. 1997.

[118] M. Soljačić, M. Ibanescu, S. G. Johnson, Y. Fink, and J. D. Joannopoulos, "Optimal bistable switching in nonlinear photonic crystals," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 66, no. 5, 2002, Art. no. 055601.

[119] B. Xu and N.-B. Ming, "Experimental observations of bistability and instability in a two-dimensional nonlinear optical superlattice," *Phys. Rev. Lett.*, vol. 71, no. 24, pp. 3959–3962, Jul. 2002.

[120] F. D.-L. Coarer *et al.*, "All-optical reservoir computing on a photonic chip using silicon-based ring resonators," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, pp. 1–8, Nov. 2018.

[121] M. Miscuglio *et al.*, "All-optical nonlinear activation function for photonic neural networks [Invited]," *Opt. Mater. Express*, vol. 8, no. 12, pp. 3851–3863, Dec. 2018.

[122] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," *Phys. Rev. X*, vol. 9, no. 2, 2019, Art. no. 021032.

[123] N. Cohen, O. Sharir, and A. Shashua, "On the expressive power of deep learning: A tensor analysis," in *Proc. Conf. Learn. Theory*, 2016, pp. 698–728.

[124] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6231–6239.

ABOUT THE AUTHORS

Qixiang Cheng (Member, IEEE) received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2014.

He joined the Shannon Laboratory, Huawei, China, where he researched future optical computing systems. From September 2016 to November 2019, he was a first Postdoctoral Researcher and then a Research Scientist with the Lightwave Research Laboratory, Columbia University, New York, NY, USA. He has been appointed as a University Lecturer in photonic devices and systems at the University of Cambridge, since January 2020. His current research interests focus on system-wide photonic integrated circuits for optical communication and optical computing applications, including a range of optical functional circuits, such as packet-, circuit-, and wavelength-level optical switch fabrics, massively parallel transceivers, optical neural networks, and optical network-on-chip.

Jihye Kwon (Student Member, IEEE) received the B.S. degree in mathematical sciences and the M.S. degree in computer science and engineering from Seoul National University, Seoul, South Korea, in 2012 and 2014, respectively. Currently, she is working toward the Ph.D. degree in computer science at Columbia University, New York, NY, USA.

She was an Intern at the IBM T. J. Watson Research Center, Ossining, NY, USA, from 2015 to 2017. Her research interest includes system-level design methodology enhanced via learning and interaction, high-level synthesis for designing hardware accelerators, real-time scheduling theory, and solving optimization problems.

Miss Kwon received the Presidential Fellowship from Columbia University, during her Ph.D. studies.

Madeleine Glick (Senior Member, IEEE) received the Ph.D. degree in physics from Columbia University, New York, NY, USA, for research on electro-optic effects of GaAs/AlGaAs quantum wells.

From 1992 to 1996, she was a Research Associate with CERN, Geneva, Switzerland, as a part of the Lightwave Links for Analogue Signal Transfer Project for the Large Hadron Collider. From 2002 to 2011, she was a Principal Engineer with Intel Research Cambridge, U.K., Intel Research Pittsburgh, Pittsburgh, PA, USA, leading the research on optical interconnects for computer systems. She is currently working in optical networking with Columbia University. She joined the Department of Physics, École Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, where she continued her research in electro-optic effects in GaAs and InP-based materials. Her research interests are in applying photonic devices and interconnects to computing systems.

Dr. Glick is currently a Fellow of the Institute of Physics and a Senior Member of OSA.

Meisam Bahadori received the B.Sc. degree (Honors) in electrical engineering, majoring in communication systems, and the M.Sc. degree (Honors) in electrical engineering, majoring in microwaves and optics, from the Sharif University of Technology, Tehran, Iran, in 2011 and June 2013, respectively, and the Ph.D. degree in electrical engineering from the Lightwave Research Laboratory, Columbia University, New York, NY, USA, in 2018, with a focus on silicon photonics.

From fall 2011 to spring 2014, he worked as a Research Assistant with the Integrated Photonics Laboratory, Sharif University of Technology. He joined with the Lightwave Research Laboratory, Columbia University, in fall 2014. His current research interests include silicon photonic devices, thin-film lithium niobate photonics, and nanophotonics.

Luca P. Carloni (Fellow, IEEE) received the Laurea degree *(summa cum laude)* in electrical engineering from the Universita' di Bologna, Bologna, Italy, in 1995, and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, USA, in 1997 and 2004, respectively.

He is currently a Professor of computer science with Columbia University, New York, NY, USA. He has authored over 130 publications and holds two patents. His current research interests include system-on-chip platforms, system-level design, distributed embedded systems, and high-performance computer systems.

Dr. Carloni is a Senior Member of the Association for Computing Machinery. He was selected as an Alfred P. Sloan Research Fellow in 2008. He was a recipient of the Demetri Angelakos Memorial Achievement Award in 2002, the Faculty Early Career Development (CAREER) Award from the National Science Foundation in 2006, the ONR Young Investigator Award in 2010, and the IEEE CEDA Early Career Award in 2012. His paper on the latency-insensitive design methodology was selected for the Best of ICCAD in 1999, a collection of the best papers published in the first 20 years of the IEEE International Conference on Computer-Aided Design. In 2013, he was the General Chair of Embedded Systems Week, the premier event covering all aspects of embedded systems and software.

Keren Bergman (Fellow, IEEE) received the B.S. degree from Bucknell University, Lewisburg, PA, USA, in 1988, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1991 and 1994, respectively, all in electrical engineering.

She is currently the Charles Bachelor Professor with Columbia University, New York, NY, USA, where she also directs the Lightwave Research Laboratory. She leads multiple research programs on optical interconnection networks for advanced computing systems, data centers, optical packet switched routers, and chip multiprocessor nanophotonic networks-on-chip.

Dr. Bergman is a Fellow of OSA.