InP photonic integrated multi-layer neural networks

APL Photonics

InP photonic integrated multi-layer neural networks: Architecture and performance analysis

Cite as: APL Photonics **7**, 010801 (2022); https://doi.org/10.1063/5.0066350
Submitted: 10 August 2021 • Accepted: 20 December 2021 • Accepted Manuscript Online: 20 December 2021 • Published Online: 13 January 2022

Bin Shi, Nicola Calabretta and Ripalta Stabile

AIP Publishing

InP photonic integrated multi-layer neural networks: Architecture and performance analysis

Cite as: APL Photon. 7, 010801 (2022); doi: 10.1063/5.0066350
Submitted: 10 August 2021 • Accepted: 20 December 2021 •
Published Online: 13 January 2022

Bin Shi,[a] Nicola Calabretta, and Ripalta Stabile

AFFILIATIONS

Institute for Photonic Integration, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands

Note: This paper is part of the APL Photonics Special Topic on Photonics and AI in Information Technologies.
[a] **Author to whom correspondence should be addressed:** b.shi1@tue.nl

ABSTRACT

We demonstrate the use of a wavelength converter, based on cross-gain modulation in a semiconductor optical amplifier (SOA), as a nonlinear function co-integrated within an all-optical neuron realized with SOA and wavelength-division multiplexing technology. We investigate the impact of fully monolithically integrated linear and nonlinear functions on the all-optical neuron output with respect to the number of synapses/neuron and data rate. Results suggest that the number of inputs can scale up to 64 while guaranteeing a large input power dynamic range of 36 dB with neglectable error introduction. We also investigate the performance of its nonlinear transfer function by tuning the total input power and data rate: The monolithically integrated neuron performs about 10% better in accuracy than the corresponding hybrid device for the same data rate. These all-optical neurons are then used to simulate a 64:64:10 two-layer photonic deep neural network for handwritten digit classification, which shows an 89.5% best-case accuracy at 10 GS/s. Moreover, we analyze the energy consumption for synaptic operation, considering the full end-to-end system, which includes the transceivers, the optical neural network, and the electrical control part. This investigation shows that when the number of synapses/neuron is >18, the energy per operation is <20 pJ (6 times higher than when considering only the optical engine). The computation speed of this two-layer all-optical neural network system is 47 TMAC/s, 2.5 times faster than state-of-the-art graphics processing units, while the energy efficiency is 12 pJ/MAC, 2 times better. This result underlines the importance of scaling photonic integrated neural networks on chip.

I. INTRODUCTION

Massive volume of data demands wider capacity and higher speed of information processing. The extraction of effective information from databases remains a challenge as it requires huge power and processing time. The new computing paradigm of non-von-Neumann architectures has begun to unfold,[1] leading to the development of large neuromorphic machines that now exceed the energy and size-efficiency walls of classical platforms,[2–8] because of their inherent parallel computational schemes. These deployments are mainly based on the spiking architectural model[9] that very recently have shown the potential to outperform multi-layer perceptron (MLP) models.[10] Nevertheless, being more complex, these models are still not fully understood, unlike the more advanced Deep Learning (DL) models. The rich DL model portfolio can be indeed utilized in digital graphics processing unit (GPU) and tensor processing unit (TPU) engines as well as in the constantly growing number of emerging artificial neural network (ANN)-based analog electronic AI chipsets: Mythic's architecture,[11] for example, can yield high accuracy inference applications within a remarkable energy efficiency of just 0.5 pJ/MAC. However, the size and energy advantages of electronic processing elements are naturally counteracted by the speed and power limits of the electronic interconnects inside the circuits due to RC parasitic effects, with current machines hardly exceeding GHz clock frequencies, exacerbating power dissipation issues, and limiting the achievable data throughput.[12]

Neuromorphic approach has been applied to optical computing: In contrast to electronics, there is negligible energy overhead for moving light encoded information around, which enables unprecedented circuit interconnectivity and speed. Moreover, bit-rate agnostic photonics has the potential to enable higher bandwidth applications. A number of photonic accelerators have been proposed based on discrete optical components and micro-optics as well as on photonic integrated devices.[13–15] This emerging technology is capable of producing high processing bandwidths with high power efficiency.[16] The large parallelism, energy efficiency, and ease of broadcast/multicast capabilities of photonics are well suited for the design of highly efficient and scalable neural network accelerators. By exploiting the properties of photonics, linear transformations can efficiently be performed at high data rates without consuming significant power.[17,18] The advantages of the parallel nature of light are now being exploited via coherent electrical field summation[19,20] and wavelength-division multiplexing (WDM) optical power addition based photonic integrated networks;[21–24] however, crosstalk, noise accumulation, and low dynamic range prevent further scalability, even when using phase change materials for zero-electrical power computation.[23]

Recently, we have proposed a new deep neural network (DNN) architecture that exploits indium phosphide based photonic integrated circuits.[24,25] By setting the gain of the semiconductor optical amplifier (SOA) as the (trained) weighted factor to the WDM input, the cross-connect is used as an analog engine with off-line nonlinear functions. Feeding the layer output back to the optical input and reconfiguring the on-chip weight matrix, a feed-forward photonic neural network is demonstrated.[25] A linear synaptic function, also called weighted addition, and a nonlinear function, known as activation function, are the base functions of an artificial neuron. Photonic integrated linear[19,26,27] and nonlinear functions[26,28] have been recently demonstrated, relying on hybrid integration schemes or involving electro-optical conversions, preventing further scalability of photonic neural networks.

In this paper, we analyze the performance of a deep neural network architecture concept based on the use of photonic cross-connects, where the combination of space and wavelength selection is exploited to implement, respectively, the axon terminals and the synaptic operations in a photonic artificial neural network. After the description of the overall computational architecture in Sec. II, the co-integration of SOA-based synaptic operations [in the form of a combination of SOAs and array waveguide gratings (AWGs)] and nonlinearity [in the form of a fully integrated wavelength converter (WC) based on cross-gain modulation (XGM)] is studied in Sec. III to enable a fully monolithically integrated all-optical neuron and therefore an all-optical neural network (AONN). In Sec. IV, we simulate the all-optical network to solve the handwriting digit classification problem to evaluate final accuracy. Finally, in Sec. V, we analyze the energy consumption of the complete end-to-end system.

II. ALL-OPTICAL SOA-BASED DNN

The overall envisaged all-optical deep neural network scheme based on the use of the cross-connect circuitry is depicted in Fig. 1. The wavelength division multiplexed (WDM) signal from N input neurons (one wavelength from each input neuron) is fan-out toward the following M neurons of the first hidden layer. At each ith neuron of this layer (highlighted through an orange box), the multiple-wavelength signal is demultiplexed into N signals, which are multiplied, via an SOA, by the weight $w_{i,j,k}$ of the ith neuron from the jth axon (λ_j) and in the kth layer. The weighted signals, being encoded in different colors, are then summed up via an AWG-based multiplexer. This first circuitry (black dashed box in Fig. 1) corresponds to the linear part of the ith neuron, whose output is sent out to a nonlinear function block $NL_{i,k}$ of the same neuron of the kth hidden layer. This is realized via an SOA (red dashed box in Fig. 1), where the enabled XGM is used to output a wavelength-converted light, modulated by the total power of all WDM channels at the input of the SOA-based wavelength converter (SOA-WC), for the conversion into a different single wavelength, λ_i', which represents the output of this neuron. The outputs from all the neurons of the kth hidden layer are then combined and broadcast again toward all the neurons of the next hidden layer, and so on and so forth. It is important to note that the shuffle network here is obtained by combining AWGs and one big 1:M splitter (for example, moving from the first to the second hidden layer), in place of M times AWGs, which would deteriorate crosstalk as well as introduce a deleterious path dependent loss.

Here, the SOA technology is exploited in combination with the AWG technology for multiple reasons: The optical amplifiers are employed for setting the weight matrix and providing on-chip gain for scalability, while the AWGs are used to filter out the out-of-band noise built up by cascading multiple stages of SOAs in

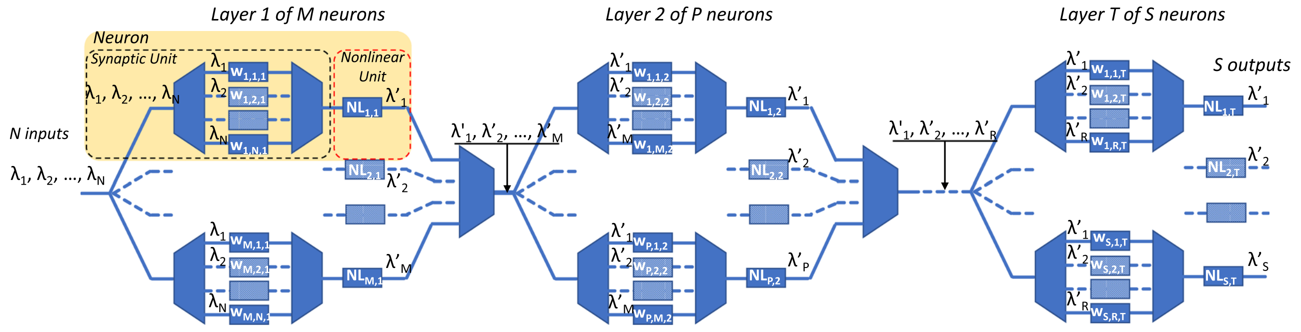

FIG. 1. Schematic of the foreseen deep neural network based on SOA-based weight and non-linear functions. NL: non-linear function. $w_{i,j,k}$ = weight of the ith neuron, of the jth axon (λ_j), and of the kth layer.

order to increase the weight resolution as well as to carry out the needed multiplexing and demultiplexing functions. In Ref. 25, we have demonstrated the synaptic operation using an 8 × 8 InP SOA-based cross-connect chip, followed by an array of photodetectors, to further process the signals in the electrical domain and to perform an analysis of the sources of error. Indeed, a reduction in accuracy happened, which was dominated by the electro-optical conversions needed to move from the optical linear function to the electrical nonlinear function as well to progress from one layer on chip to the next one, which suggested moving to an all-optical approach.[25]

We now investigate other sources of errors and scalability properties of the linear neuron, specifically the crosstalk as a function of the number of channel inputs (axons). The optical crosstalk, coming from the AWGs, limits the linear circuitry scalability as soon as the number of neuron inputs (channels) increases. For this reason, we analyze the normalized root mean square error (NRMSE) of the synaptic unit of the neuron, as shown in Fig. 1 (black dashed box), while tuning the total input power and the total number of neuron inputs for a channel spacing of 100 GHz. This has been analyzed via the VPIphotonic Design Suite (using parameters as detailed in Ref. 29). In Fig. 2(a), the colored lines represent the error obtained after the synaptic unit for 4, 8, 16, 32, and 64 inputs/neuron while tuning the total input power of the WDM input to the neuron from −25 to 20 dBm. The results show that there is an optimized optical power operation point for reaching the minimum error. It is notable how this point shifts down right side for a larger number of neuron inputs. To better visualize and explain this trend, the same NRMSE is plotted as a function of the number of the input channels for a fixed input power of 5 dBm at the AWG input [see Fig. 2(b)]. When increasing the number of neuron inputs, the error decreases, as shown in blue line, while it starts slightly increasing only for a number of channels higher than 32: The vertical scalability of the neural network (height), and therefore a higher channel number, results in an increase in the resolution of the linear summation output, since more channels at the input contribute to increasing the total number of the output signal levels within the same dynamic range, resulting in a smoother output signal pattern. In particular, the error is found to increase for 64 channel inputs due to the limited modeled SOA bandwidth (71.5 nm 3-dB gain bandwidth); in fact, 64 channels spaced 100 GHz already fill up 51.2 nm bandwidth. The red line in Fig. 2(b) plots the input power dynamic range (IPDR) as a function of the number of the input channels per neuron and for an NMRSE <0.09: for this level of error, we have previously shown that a three-layer neural network results in <5% degradation of the prediction accuracy for an image classification problem.[29] The IPDR increases from 25 to 36 dB, which is partly attributable to the large SOA linear regime (−5 dBm input saturation power) but also to the fact that with the increasing number of input channels, the power fed to the individual weight SOA will be much lower than the input saturation power, making the SOAs working in the linear regime for a wider input power range. The trend slows down when the number of channels approaches 64 since we come closer to the bandwidth edges of the SOA.

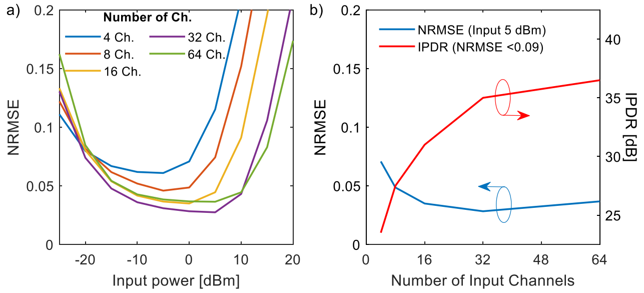

FIG. 2. (a) Error obtained when tuning the total input power to the neuron from −25 to 20 dBm per different input channel numbers. (b) Error variation (5 dBm input power, blue line) and IPDR (NMRSE <0.09, red line) when changing the numbers of input channels from 4 to 64 channels. The AWG channel spacing is set at 100 GHz.

III. SOA-BASED INTEGRATED ALL-OPTICAL NEURON

So far, we have proposed to use SOA and AWG to implement optical linear neurons.[24] In this section, we investigate the possibility of realizing an all-optical SOA-based neuron to realize multi-layer neural networks and avoid electro-optical conversions for improving energy efficiency while still guaranteeing a good accuracy. To this aim, the optical output of the synaptic operation is input straight to an SOA-based nonlinear function. The exploitation of SOA-based circuits for both the linear and nonlinear functions of an artificial photonic neuron enables the monolithic integration of both functionalities to overcome optical loss issues deriving from a hybrid approach. We first study the nonlinear function based on a wavelength converter (Sec. III A), and then we investigate the overall performance of the complete neuron, integrating the optical linear neuron with the SOA-WC optical nonlinear function (Sec. III B).

Before describing the experimental measurements and simulations in Subsections III A and III B, it is important to discuss the assumptions made on any four-wave mixing (FWM) effect happening within the nonlinear SOA. Depending on the wavelength separation, input power, and the number of WDM channels, the FWM inside the SOA may have a non-negligible influence on the overall performance. However, this is not considered in the simulation, neither is observed in the experimental phase. In fact, this effect is neglectable when the detuning between the probe channel and the pump frequency $\Delta f \gg 1/(2\pi \cdot \tau)$, where τ is the carrier lifetime of the used SOA. In this work, the carrier lifetime is estimated to be 200 ps in the worse case,[30] and the channel spacing at the input is 100 and 400 GHz for the simulations and the experimental work, respectively, which results in detuning that is far greater than $1/(2\pi \cdot \tau)$ $\approx$ 1 GHz. By exploiting the methods in Refs. 31 and 32, we estimate that the conjugate signal generated by the FWM effect has a power of the order of <−64 dBm when the detuning used in simulation is >100 GHz, which is even lower than the spontaneous emission noise at the neuron output. Moreover, in order to suppress the FWM for larger number of input channels, we control the total input power of the neuron by defining an appropriate scaling of the neural network. In our approach, the network size is considered scaling up with input channels N and with the same number of neurons M. In this way, the total input power to the neuron will stay constant when scaling N and M, i.e., for each channel power p_0, the total input power at the layer input $N \cdot p_0$ will be split toward M neurons as $N \cdot p_0/M$. When setting $N = M$, the total power to each neuron (yellow box in Fig. 1) will be p_0, and the power for each channel will be p_0/N. For such a condition, the FWM effect results reduced due to the decrease in the input signal power as well as to further detuning of the individual input channels. Finally, using unequal channel spacing at the WDM inputs,[33,34] the FWM effect can be further eliminated.

The experimental setup used for the SOA-based all-optical neuron investigation is depicted in Fig. 3(a) with a micrograph of the fabricated chip shown in Fig. 3(b). A four-channel WDM optical input is composed of signal wavelengths set at 1544.0, 1546.0, 1551.0, and 1554.0 nm in order to match the nominal 3.2 nm channel separation of the on-chip AWGs with a 3-dB bandwidth of 0.8 nm and to maximize each channel optical power output. The input is modulated with pseudorandom binary sequence (PRBS) on–off keyed (OOK) data, generated by the arbitrary waveform generator (Tektronix, AWG7122B), and sent to the input of the integrated all-optical neuron after de-correlation. The WDM optical inputs are equalized and set at −12 dBm power per channel.

Inside the neuron in Fig. 3(b), the inputs are amplified with a booster SOA, which is utilized to optimize the total power at the input of the weighting SOAs. Then, the signals are weighted with individual SOAs after channel demultiplexing by the AWG and combined again with an AWG-based multiplexer and fed to the SOA-WC based optical nonlinear function, whose pump laser is fully integrated on chip. This provides a converted output at 1549.0 nm, chosen to be close to the center of the WDM channel bandwidth for optimizing the wavelength conversion.

The weighting SOAs are controlled by a weighting current controller (Thorlabs, MLC8200CG), with 50 μA resolution, to provide the weights in 10-bit precision, which exceeds the required precision for image classification. The current synapse control is envisioned to be realized by means of an field-programmable gate array (FPGA) controlling multi-channel current drivers[35] when further scaling the number of neuron synapses. In the future, the parallel development of ultra-compact driver ICs, of new electronic interface techniques, and of cleaver electrical control schemes seems to be a viable route toward enabling control of larger size photonic networks on chip.

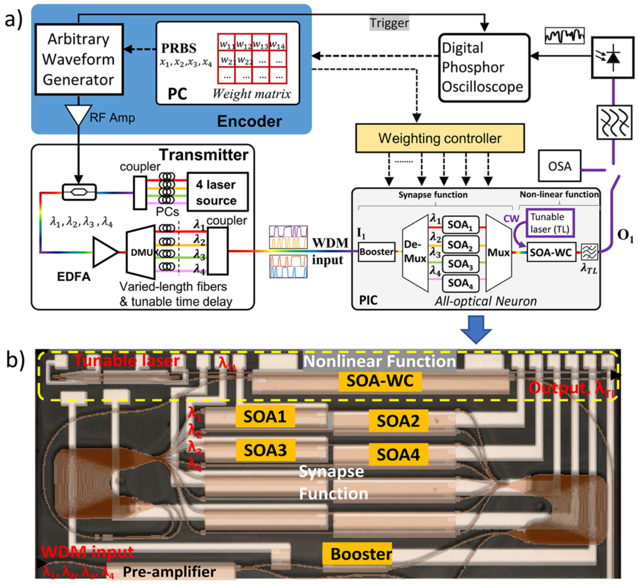

FIG. 3. (a) Experimental setup for SOA-based all-optical neuron investigation. (b) Micrograph of the integrated all-optical neuron.

The individual weight SOAs are calibrated to compensate for the wavelength conversion non-uniformity among the different channels: This calibration happens prior to the assignments of the actual weighting factors. The noise figure and the saturation output power of these SOAs are 7 dB and 8 dBm, respectively. The output of the all-optical neuron is detected by a linear avalanche photodetector (PD) and the time trace is recorded by a digital phosphor oscilloscope. The performance of the neuron is again evaluated by calculating the NRMSE between the recorded and the expected time traces at the output of the NL function, calculated using the reference pre-recorded inputs. The synaptic operation of the neuron can be expressed as a weighted addition of parallel inputs: $y = \sum w_i x_i$, where w_i is the ith weight element for input x_i, and the final output of the neuron is $o = \varphi(y)$, where φ is the nonlinear transfer function of the SOA-WC.

A. Integrated SOA-based non-linear function

The wavelength converter is the nonlinear device that we exploit as an optical nonlinear function within the neuron. The SOA-WC is also integrated on the InP platform, with an on-chip tunable laser.[36] The integration of the all-optical nonlinear function allows us to demonstrate a monolithically integrated SOA-based all-optical neuron.[37] In order to measure only the transfer function of the nonlinear part working at first as a simple inverter, we record the PRBS OOK input of the neuron and the output of the SOA-WC. The correlation map of the two is the nonlinear transfer function (NL-TF), which we can use to calculate the expected output for the entire all-optical neuron. The blue line time trace in Fig. 4(a) plots the pre-recorded 2 Gbit/s single channel input signal. Figure 4(b) presents the detected output of the SOA-WC based NL-TF in the blue line and the expected inverted signal [calculated from Fig. 4(a), with the linear transform as reference—Lin. Ref.] in the red line, resulting in an error of 0.14. By plotting the correlation map between the input and the output of the integrated SOA-WC detected at the PD, the optical nonlinear transfer function is illustrated in Fig. 4(d), where the blue crosses are the data, the black line is the linear transform, and the red line is the third-order polynomial fitted nonlinear transform. The nonlinear function shape is mainly due to the contribution of the nonlinear response of the SOA used as the wavelength converter when the booster works in transparency, with a current density of 1 kA/cm^2 and a weighting current density at 3 kA/cm^2 on average (linear regime). Then, the same nonlinear function shape is utilized as a nonlinear reference (Nonlin. Ref.) to calculate the real expectation of the output, as shown in Fig. 4(c), resulting in a smaller error of 0.08.

The pump input power of the wavelength converter can be tuned by increasing the current of the booster: A different level of pump input power provides a different transfer function shape. Figure 5(a) presents the nonlinear functions when the current density of the booster is set at 0.5, 1.0, 1.5, and 2.0 kA/cm^2, with lines in blue, red, yellow, and purple, respectively. The outputs near to

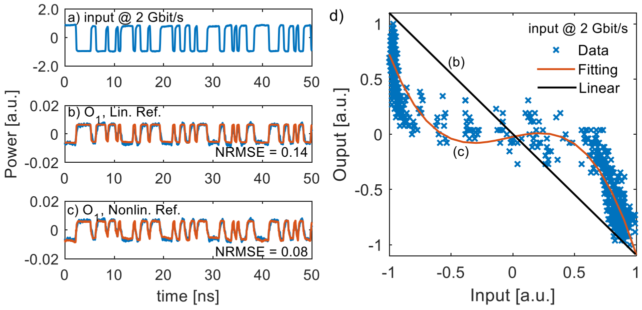

FIG. 4. (a) The input data in channel 1 at 2 Gbit/s. (b) The recorded output (blue) comparing to the expectation (red) with linear transformation as reference at the SOA-WC. (c) The recorded output (blue) and the expected output (red) with nonlinear referenced calculation. (d) The nonlinear transfer function of the SOA-WC, with correlation mapping of the input to output, with linear transformation (black) and nonlinear transformation (red).

level "−1" (for input level "1") tend to saturate when increasing the booster current because of the nonlinearity changes due to the increased input probe power to the SOA-WC and because of the nonlinearity contributed from the booster SOA itself. This confirms that we can tailor the nonlinear transfer function by acting on the booster current. We also explore the SOA-WC based NL-TF shape as a function of the data rate: Fig. 5(b) plots the nonlinear function when the input data rate is 2, 4, 6, 8, and 10 Gbit/s, with the blue, red, yellow, purple, and green line, respectively, when the booster is at 1 kA/cm^2. The shape of the nonlinear function changes only slightly when increasing the input data rate. We then translate these findings into performance metrics of the optical nonlinear function by calculating the NRMSE with respect to different shapes of the nonlinear function. Figure 5(c) plots the error variations of the output of the NL-TF when tuning the injection current density of the booster SOA from 0 to 2.5 kA/cm^2, with the blue line obtained when considering the neuron output as the linear inverted output (with linear transform as reference) and the red line when considering the nonlinear transform as reference. The booster SOA is operated in the linear regime to minimize the nonlinearities introduced at the weighting element inputs, since the overall weighted addition operation is meant to be a linear operation. By changing the booster SOA current, we can find the optimal operation point for minimized error induced by the nonlinear function, in this case corresponding to a current of 1 kA/cm^2 [Fig. 5(c)]. The noticeable offset between the blue and red curves indicates that the nonlinearity of the SOA-WC has quite some effect on the error reduction. Figure 5(d) plots the error variation when changing the data rate of the input from 2 to 10 Gbit/s, with the blue line showing the error related to the linear transform reference and the red line showing the error related to the nonlinear transform reference. In both cases, the error of the nonlinear function increases with the input data rate. Again, the nonlinear function improves accuracy, moving from 0.08 to 0.15 NRMSE when increasing the data rate up to 10 Gbit/s. The deterioration in accuracy for the higher data-rate is mainly due to the limited carrier lifetime of the integrated SOA-WC, which cannot fully follow the speed of the incoming optical signal. The offsets between the blue and red lines in both Figs. 5(c) and 5(d) show that the use of the correct nonlinear transfer function reduces the error of up to 50%, compared to the case when we use the SOA-WC with its simply linear response.

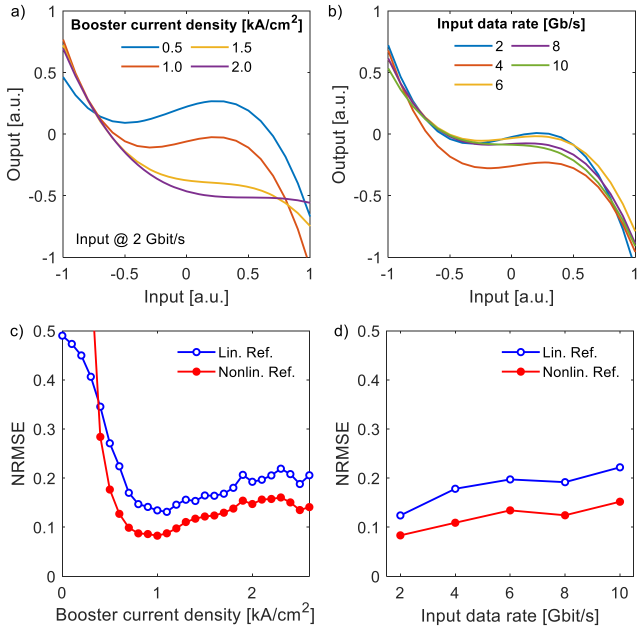

FIG. 5. (a) The nonlinear transfer function when the booster current density is set at different levels. (b) The nonlinear transfer function when the input data rate changes from 2 to 10 Gbit/s. Error obtained at the neuron output when tuning the booster current density (c) and when tuning the input data rate (d), comparing to expectation calculated using linear transform as reference (blue) and nonlinear transform as reference (red).

B. All-optical monolithically integrated neuron

The monolithic integration of the synaptic operation and the optical nonlinear function allows us to investigate the performance of the SOA-based all-optical neuron concept. The four-channel WDM PRBS-OOK input is coupled at the neuron input, with a data rate of up to 10 Gbit/s. The output is detected and compared to the calculated time trace with the NL-TF obtained following the procedure explained in session III-A. Figure 6(a) plots the time traces as a linear combination of the weighted input data, where the red line presents the recorded signal and the blue line is the expected linear combination of the weighted addition, without NL-fitting, resulting in an error of 0.17. Figure 6(b) instead shows the output of the recorded output signal with the nonlinear transform reference, where the blue line is the recorded signal and the red line is the expectation, resulting in a smaller error of 0.15—a 10% error reduction. We also change the number of input channels and tune the data rate of the input signals to better analyze the performance of this all-optical neuron. Figure 6(c) illustrates the error of the complete optical neuron output. The blue circle, red triangle, and yellow square symbols represent the errors of the all-optical neuron when the input channel changes from 1, 2 to 4 channels, respectively. In line with Fig. 5(d), the curves show that the output error increases

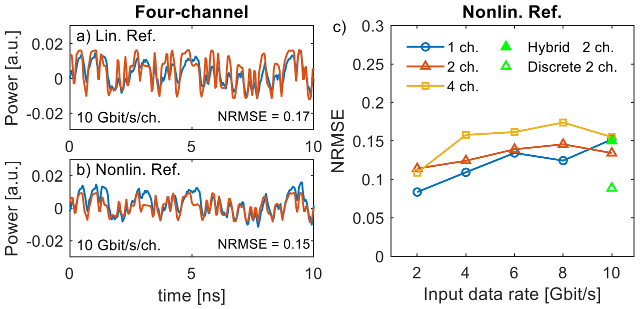

FIG. 6. The four-channel weighted addition recorded output (blue) comparing to the expected (red) (a) linear transformed reference, without NL-fitting, and (b) nonlinear transformed reference with NL-fitting. (c) The error obtained at the neuron output for 1, 2, and 4 channel weighted addition (blue, red, and yellow) when tuning the data rate from 2 to 10 Gbit/s per channel, with green filled triangle denoting the error of the two-channel hybrid integrated neuron and the unfilled triangle of denoting the error of the two-channel discrete neuron with optimized SOA-WC.

with the input data rate. Moreover, with the increase in the channel number, the error tends to increase as well. With more channel input to the SOA-WC, the nonlinearity at the SOA-WC is reduced as the increasing input probe power will push the cross-gain modulation regime toward a linear conversion regime. This means that an optimization of the operation regime of the wavelength converter is needed to help increase its nonlinearity, e.g., by tuning the power of the CW laser. Moreover, in Fig. 6(c), we also add a green-filled triangle to show an average error of 0.15, which is obtained when combining the integrated linear unit with a discrete nonlinear SOA wavelength converter,[38] with 10 Gbit/s per channel input and two channel weighted addition. This shows that the monolithically integrated all-optical neuron performs 10% better in terms of error introduction than the hybrid case under the same data rate condition. One reason for that can be that a discrete implementation generates additional noise due to the off-chip amplification. Finally, the integration of the tunable laser and SOA-WC also reduces the total power consumption as the external laser is not required, neither the additional off-chip amplifier. Further investigation shows that by using discrete SOA-WCs with optimized carrier dynamics, the multi-level conversion brings to a calculation error less than 0.09,[39] shown as green-unfilled triangles in Fig. 6(c). In Sec. IV, we show the simulation of an all-optical multi-layer neural network by exploiting both the synaptic operation and the nonlinear function as realized and measured so far.

IV. MNIST DATASET CLASSIFICATION WITH AN SOA-BASED ALL-OPTICAL NEURAL NETWORK

The combination of the linear neuron with the wavelength converter (Sec. III) eventually converts the multiple weighted wavelength inputs, after their addition, into one single wavelength which is the actual output of the complete neuron (yellow box in Fig. 1). In particular, the recorded transfer function of the integrated SOA-based wavelength converter, shown in Fig. 4(d), has been evaluated in an analog manner, with the power summation at the input of the SOA-WC being a multi-level signal. Therefore, the same transfer function will also work with multi-level WDM signals. This nonlinear function is then used to train the neural network on the computer via TensorFlow,[40] while the pre-trained weighted matrix can be applied to the all-optical neural network to run inference and evaluate the accuracy.

The handwritten digit classification problem[41] with modified National Institute of Standards and Technology (MNIST) dataset is one of the benchmarking problems used for the performance appraisal of a neural network. The MNIST dataset contains 60 000 training samples and 10 000 testing samples and includes ten categories of digits from 0 to 9. In Sec. II, we have discussed that the linear synaptic operation of the SOA-based neuron can allow more than 64 channel inputs, with the introduction of negligible error. Here, we indeed simulate the all-optical neural network with input layer neurons with 64 channel inputs each. To encode the input image into 64 channels by means of multi-level modulation with 9-bit resolution, we preprocess the images in the dataset to reduce their resolution from 28 × 28 to 8 × 8 pixels. Figure 7(a) illustrates the data preprocessing for the input of the neural network (NN). The 256 level gray data are first converted into a black and white image with a threshold at level 128 and cropped into 24 × 24 pixels at the center. The images are then converted to 8 × 8 pixels with every 3 × 3 pixels encoded into 512 grayscale levels, i.e., 9 bits-resolution. For solving this digit classification, a two-layer NN is structured as shown in Fig. 7(b). On the first layer, there are 64 neurons where each of the weighted addition output is followed by the optical nonlinear function obtained in Sec. III, and on the second (output) layer, ten linear neurons are used to represent the ten digits, from 0 to 9. In the optical neural network (ONN) implementation, the inputs and the weights are usually normalized in order to ease the optical modulation and the dynamic weighting control. This is implemented in simulation by applying batch normalization and weight normalization. To train the NN for MNIST dataset classification, the ADAM optimizer is utilized due to its fast convergence,[42] which makes the training process more efficient.

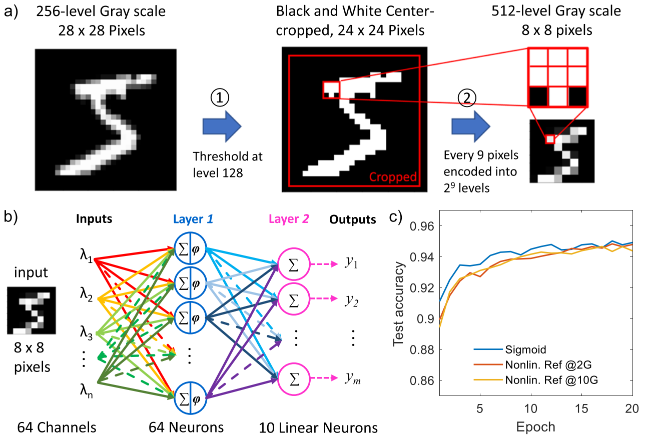

FIG. 7. (a) Preprocessing of the input MNIST handwritten images from 728 to 64 pixels. (b) The two-layer neural network structure for classifying MNIST handwritten digits. (c) The test accuracy per epoch when training the two-layer neural network with the sigmoid function (blue line) and the nonlinear functions when the data rate is 2 GS/s (red) and 10 GS/s (yellow).

We train this two-layer structure with the current third order polynomial nonlinear function without noise induction as a reference. The trained weighted matrix is then applied to the ONN model to investigate the performance of the optical network under error induction and contribution from the linear and the nonlinear units. Moreover, we benchmark this same shallow neural network for the same data in Fig. 7(a), but using the sigmoid nonlinear function. The test accuracy is recorded after every update of the weighting matrices when training the neural network in TensorFlow. Figure 7(c) presents the test accuracy as a function of the training epochs for different nonlinear functions: when the nonlinear function is the conventional mathematical sigmoid function (blue), when it corresponds to the transfer function observed at 2 GS/s per channel (orange), and when it corresponds to the transfer function obtained at 10 GS/s as input (yellow). Note that here we do not consider yet the influence of the all-optical neuron impairments. The curves show that the NN is converging after 15 epochs of training and that all considered nonlinear functions yield a similar final test accuracy of ~94.5% after training.

To take into account the error induced by the all-optical neuron, we consider the distortion contribution due to the linear part of the neuron (described in Sec. II) and the distortion contribution due to the nonlinear part of the neuron (analyzed in Sec. III). In particular, the distortions are included here as additive white Gaussian noise, assuming that the signal spontaneous emission beating noise dominates the contribution,[43] which is added after the linear output and the nonlinear output. By tuning the standard deviation of the Gaussian noise, the same error levels as the ones observed experimentally can be reproduced. The same inference is now run in the case that impairments are induced in the optical neuron: Figs. 8(a) and 8(b) illustrate the colormap of the prediction accuracy (Acc.) as a function of the noise levels of both the linear and nonlinear functions of a neuron: these are scanned from 0 to 0.5 for both 2 and 10 GS/s input per channel, respectively. The accuracy in both cases obviously decreases when increasing the error at the output of both linear and nonlinear units. The red line shapes in Figs. 8(a) and 8(b) show the expected accuracy that the AONN system will have for a measured error level ranging from 0.05 to 0.10 for the linear operation (according to the error induced with 64 channel inputs, discussed in Sec. II) and for the nonlinear errors ranging between 0.08 and 0.11 for 2 Gbit/s input and between 0.10 and 0.15 for 10 Gbit/s input, respectively, as recorded during the experiments. For these same areas, an accuracy degradation of 2%–8% and 5%–15% for 2 and 10 GS/s input, respectively, is obtained, compared to the trained accuracy of 94.5%. The elliptical shape in Fig. 8(a) is due to a different deviation of the Gaussian noise distribution on the linear and nonlinear unit, while the circular shape in Fig. 8(b) is due to a more uniform variation for both units. These suggest that with 10 GS/s input, the two-layer all-optical engine, including 64 neurons in the first layer, with 64 synapses per neuron, and 10 neurons at the second layer fully connected, can perform 4.7 × 10^{13} MAC/s, which provides ~2.5 times faster computation than the state-of-the-art GPUs[44] and the same order as the TPU,[45] considering only 5% best-case accuracy degradation and 10 GHz speed nonlinear processing, which is not available in GPUs and TPUs. Training the AONN with the addition of the estimated distortion from the linear and the nonlinear unit is expected to reduce the influence of the noise and preserve the high prediction accuracy of the NN using the wavelength converter as the nonlinear function instead of the conventional sigmoid function. In the future, we envision that the scaling to 64 input neurons in our network system can be realized by interfacing the chip with high-speed state-of-the-art transceiver modules[46] or with co-packaged optics[47] in a multi-chip package.

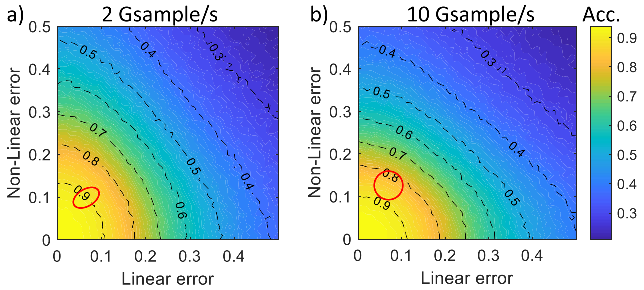

FIG. 8. The accuracy of the simulated noised ONN for the nonlinear function recorded when input at (a) 2 GS/s and (b) 10 GS/s. Red circles represent the expected performance of using our all-optical neuron in the ONN.

V. SYSTEM ENERGY CONSUMPTION ANALYSIS

In this section, we estimate the power consumption on the end-to-end (digital-to-optical-to-digital) system enabling the implementation of the optical neural network. Figure 9 shows the schema of the complete ONN system, which includes the transmitter, the optical chip, the receiver, the digital signal processor, and the control unit. The system overall is controlled by the control unit (Ctrl), which is interfaced with the computer and includes a field-programmable gate array (FPGA) and a digital signal processor (DSP). Here, we use an FPGA for the sake of fast development and reconfiguration flexibility.[48,49] However, application specific integrated circuits (ASICs) can also be used to reduce the power consumption even further.[50] To analyze the effective power consumption of the ONN, all the components in the system should be taken into account. The transmitter (Tx) includes lasers, modulators, and DACs, which are used to drive the modulators. The ONN includes the ONN chip and its control DACs and drivers for weighting. The receiver (Rx) consists of photodetectors and the corresponding ADCs.

The energy consumption of the system is analyzed by considering different operation modes of the ONN within the end-to-end

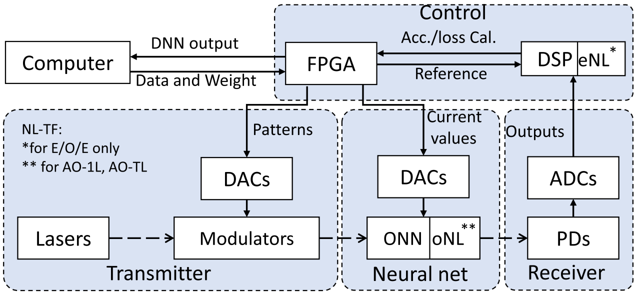

FIG. 9. Schematic of the considered end-to-end optical DNN system.

system. Here, we consider three scenarios: (1) E/O/E with one linear layer (E/O/E), (2) All-Optical with one complete layer (AO-1L), and (3) All-Optical with multiple layers (AO-TL), where T stands for T layers. In the E/O/E approach, the optical chip is used to calculate linear matrix multiplication, while the nonlinear function is realized on the DSP, with the data received at the PDs. In the all-optical approaches, the nonlinear function is co-integrated with the linear optical neuron, and the output (at each layer output for the AO-1L case or at the end of the complete multiple-layer NN for the AO-TL case) is obtained via linear PDs. A more general ONN with N-inputs M-outputs and T-layers is now analyzed, including the end-to-end system performance, for these three different operation modes. The operations executed by the ONN systems are different for these cases, depending on if a single layer or multiple layers are implemented. For the single-layer implementations, as in cases (1) and (2), the DNN needs to be decomposed into layers and analyzed layer by layer, which is not necessary in case (3) for the same network implementation.

For the inference of a trained DNN, the data and weight matrix are loaded to the FPGA via the interface with a computer. The FPGA generates the electrical patterns as well as the weight control currents, which feed to the modulator DACs and the weight DACs and drivers, respectively, as shown in Fig. 9. The electrical patterns are imprinted on the laser beams and sent to the optical neural network chip. The chip is controlled with the analog currents coming from the respective DACs and amplified at the drivers, with which the matrix multiplications are calculated. For the E/O/E case, the detected linear output is converted into digital signals by the ADCs, and then the DSP unit processes the signals executing the nonlinear transfer function. The outputs are then sent back to the FPGA, which generates the patterns for the next layer. The next layer follows the same procedure. At the output layer, the outputs of the last layer nonlinear functions will be further processed by the FPGA and compared with the reference labels to provide the final prediction, which is then passed to the computer. Therefore, the power consumption of the E/O/E single-layer system can be calculated as

$$P_{E/O/E} = N \times P_{Tx} + (N \times M) \times P_w + M \times (P_{Rx} + P_{eNL} + P_{ctrl}), \quad (1)$$

where P_{Tx} is the power of transmitter per channel, P_w is the power for each weighting, including the power of DAC and the current driver for the ONN, P_{Rx} is the power of receiver, P_{eNL} is the power for the electrical nonlinear function, and P_{ctrl} is the power of the control.

For AO-1L case, the procedure is similar to the E/O/E case, with the only difference that the nonlinear function is co-integrated on the optical chip. Therefore, the DSP does not carry out the nonlinear function calculation and only calculates the final accuracy at the output layer. Hence, the power of the AO-1L system can be calculated as

$$P_{AO-1L} = N \times P_{Tx} + (N \times M) \times P_w + M \times (P_{oNL} + P_{Rx} + P_{ctrl}), \quad (2)$$

where P_{oNL} is the power of the photonic nonlinear function. Finally, for the AO-TL case, the FPGA and DSP are not required to process and update the inputs and weights for the next layer, but the DSP will calculate the loss and accuracy based on the final outputs and the reference labels. Therefore, the power consumption of the AO-TL system can be calculated as

$$\begin{aligned} P_{AO-TL} &= N \times P_{Tx} + (N \times M \times T) \times P_w \\ &\quad + M \times T \times P_{oNL} + M \times (P_d + P_{ctrl}). \end{aligned} \quad (3)$$

The required number of components of the three different scenarios and the power values used in the system power analysis are listed in Table I. These values are considered when using state-of-the-art components that fit into the scheme of the SOA-based all-optical neural network structure as described in Sec. II.

Considering the delays related to all the components, the total time for the E/O/E system to execute one epoch can be specified as

$$\begin{aligned} t_{E/O/E} &= T \times (S_N/f_{Tx} + 1/f_{Tx} + t_{Tx} + t_{oLin} + 1/f_{Rx} + t_{Rx} \\ &\quad + S_N/f_{eIO} + t_{eNL} + t_{FPGA} + t_{e-inter}) + t_{acc}, \end{aligned} \quad (4)$$

for an AO-1L single layer system is calculated as

$$\begin{aligned} t_{AO-1L} &= T \times (S_N/f_{Tx} + 1/f_{Tx} + t_{Tx} + t_{oLin} + t_{oNL} + 1/f_{Rx} \\ &\quad + t_{Rx} + S_N/f_{eIO} + t_{FPGA} + t_{e-inter}) + t_{acc}, \end{aligned} \quad (5)$$

TABLE I. Components in the optical neural network system.

Components		E/O/E	AO-1L	AO-TL	Unit P (mW)	References
Tx	Laser	N	N	N	150	51
	Mod.	N	N	N	20	52
	DACs	N	N	N	25	53
ONN	Weight Elements	N × M	N × M	N × M × T	30	25
	DACs	N × M	N × M	N × M × T	25	53
	o-NL	-	M	M	150	37
Rx	PDs	M	M	M	5	54
	ADCs	M	M	M	25	53
	e-NL	M	-	-	200	49
Ctl	Accuracy Cal. Unit	M	M	M	200	49
	FPGA	M	M	M	200	

and for an AO-TL multi-layer system is calculated as

$$t_{AO-TL} = S_N/f_{Tx} + 1/f_{Tx} + t_{Tx} + T \times (t_{oLin} + t_{oNL}) + 1/f_{Rx} + t_{Rx} + S_N/f_{eIO} + t_{FPGA} + t_{e-inter} + t_{acc}, \quad (6)$$

where S_N is the number of samples per epoch at the input of each layer and f_{Tx} and f_{Rx} are the speed of the transmitter and receiver, respectively; t_{Tx}, t_{oLin}, t_{oNL}, t_{Rx}, t_{eNL}, *and* $t_{e\text{-}inter}$ are the time delay from the transmitter, the optical linear unit, the optical nonlinear unit, the receiver, the electrical nonlinear function, and electrical interconnection, respectively; t_{FPGA} is the computational time for the FPGA to generate the patterns and the current values for the weights; and t_{acc} is the computational time of the DSP for the accuracy calculation. The average total energy consumption for epoch can be expressed as $E_{syst} = P_{syst} \cdot t_{syst}$, where E_{syst} is the total energy consumptions for the whole neural network system per epoch, t_{syst} is the time for computing one epoch of samples, and P_{syst} is the total power of the end-to-end system, all calculated, respectively, for the three operational system cases.

The energy consumption for the optical MAC operation, i.e., the synaptic operation, depends on the number of controlled elements which provide the weights if only the optical engine is considered. Here, we use the same weighting elements, i.e., the SOAs, for which the power is 30 mW on average per weight, excluding the DACs. Therefore, for an operational input data rate of 10 GHz, the resulting power consumption for one MAC is 3.0 and 5.5 pJ/MAC if we include the weight DACs. However, this estimation misses the contribution of the transceiver, the overall electrical controller, the receiver, and the off-chip computations. Therefore, the end-to-end system power and the total computational time should be considered to obtain the real performance metrics of the optical neural network. For an N-input M-neuron T-layer DNN, the total number of MAC operations is $S_N \times M \times N \times T$. Hence, the effective energy consumption—*effective* as we now include the end-to-end system overall contribution—per MAC operation is the total power of the specific end-to-end system times the total time to execute one epoch over the total number of MAC operations,

$$E_{MAC-eff} = P_{syst} \times t_{syst}/(S_N \times M \times N \times T). \quad (7)$$

The delays and computational speed for different components are listed in Table II. The values used in the calculations are considered based on off-the-shelf components. In particular, all optical delays are obtained from the actual path length, while all the electrical delays are related to the processing clock time of the off the-shelf electronics.

TABLE II. Computing time of the components in the system.

Symbol	Description	Value	Unit
N	Max. synapsis number in a neuron	64	
M	Max. neuron number per layer	64	
T	Max. layer number	10	
S_N	Input sample number	10^4	
$f_{Tx/Rx}$	Speed of the optical transmitter/receiver	10	GHz
t_{Tx}	Time delay, transmitter	5	ps
t_{oLin}	Time delay, optical linear unit	10	ps
t_{oNL}	Time delay, optical nonlinear unit	20	ps
t_{Rx}	Time delay, receiver	2	ns
t_{eNL}	Time delay, electrical NL unit	3	ns
$t_{e\text{-}inter}$	Time delay, electrical connection	100	ns
f_{eIO}	Speed, I/O connection, FPGA	10	GHz
t_{FPGA}	Time, signal processing of FPGA	3	ns
t_{acc}	Time, acc./loss calculation of FPGA	6	ns

We first investigate the size scaling of the optical neural network. As mentioned in Sec. II, the network is considered to be scaling up with $M = N$, i.e., this energy analysis is done with respect to a quadratic scaling of the network. When the increasing number of neurons M, the splitting loss will increase. As a consequence, we compensate these losses with additional laser power by increasing N, the input channel number. From Table II, it is clear that the largest DNN that we will investigate is an $M \times N \times T$ DNN with a maximum number of 64 input ×64 neuron/layer × 10 layers. Figure 10 illustrates $E_{MAC\text{-}eff}$ obtained from Eqs. (1)–(7) for different system modes of operation and looking at different parameters. Figure 10(a) illustrates the energy consumption per MAC operation when increasing the number of synapses per neuron, N, with the layer number $T = 10$ (solid lines) and increasing the layer numbers T when fixing $M = N = 64$ (dashed lines). $E_{MAC\text{-}eff}$ for the multi-layer DNN is inversely proportional to the number of synapses for all the cases. The $E_{MAC\text{-}eff}$ for E/O/E (in blue) and AO-1L (in red) are very close, as in both cases the FPGA and transmitter for the signal processing and pattern regeneration, respectively, notably increase

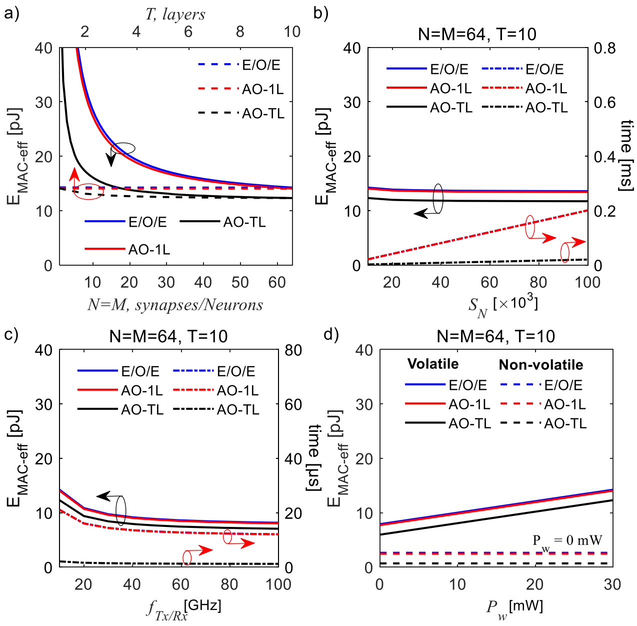

FIG. 10. The system energy consumption per MAC, $E_{MAC\text{-}eff}$. (a) $E_{MAC\text{-}eff}$ obtained when changing the synapses number N (solid) and layer number T (dashed); (b) $E_{MAC\text{-}eff}$ (solid) and total computing time (dashed) vs input sample numbers S_N; (c) $E_{MAC\text{-}eff}$ (solid) and total computing time (dashed) vs speed of transceiver; (d) $E_{MAC\text{-}eff}$ calculated (solid) and total computing time (dashed-dotted) when changing the power of weighting elements P_w (for the E/O/E (blue), AO-1L (red), and AO-TL(black) neural network).

the power consumption as well as computing time after each layer. $E_{MAC\text{-}eff}$ tends gradually to the asymptotic value of 14 pJ/MAC. The lower limit of energy consumption is set by the power consumption at the transmitter side and at the weighting elements (this power relates to the weight unit power, and therefore, it does not depend on the synapses number). For the AO-TL neural network system, avoiding the electronics to optics to electronics conversions when moving layer by layer, the computing time gets reduced considerably: The rate of change of $E_{MAC\text{-}eff}$ is faster than for the E/O/E and AO-1L cases and reaches 12 pJ/MAC for a number of 64 synapses/neuron. If the FPGA was replaced with an ASIC with optimized designs to reduce the power consumption, the effective energy consumption would have not been changed dramatically since, in these particular large-scale network systems, the elements for the control of the weight represent the main contribution. Always in Fig. 10(a), we observe that the number of synapses per neuron in the system with single layer implementations should be greater than 20 for case (1) and greater than 18 for case (2) in order to guarantee $E_{MAC\text{-}eff}$ down to 20 pJ/MAC, while for the AO-TL neural network, this value is only 6. On the other hand, the dashed lines plot $E_{MAC\text{-}eff}$ with respect to the number of layers when the synapses number N is 64. $E_{MAC\text{-}eff}$ is, in general, very flat for all three cases. The difference among the single layer cases, E/O/E (1) and AO-1L (2), with the multi-layer case, AO-TL (3), is set by the synapse number: a bigger difference is expected for a smaller synapse number.

All the graphs in Fig. 10(a) tend to an asymptotic value because the lower limit is bound to the energy consumption on each synapse control component for $M = N > 64$ and $T > 10$. Hence, we carry out all the other investigations for $M = N = 64$ and $T = 10$ while changing other parameters, such as the input sample number S_N, the speed of the transceivers $f_{Tx/Rx}$, and the power of the weighting elements P_w. Figure 10(b) presents $E_{MAC\text{-}eff}$ and the total computing time when changing input sample numbers. The power efficiency only slightly decreases with varying the input sample numbers from 10 to 100 k (solid lines) for two reasons: $E_{MAC\text{-}eff}$ is calculated on each MAC operation of each sample and the total processing time for computing (dashed lines) increases linearly from 20 to 200 μs for the single-layer E/O/E and AO-1L neural network. The computing time for the AO-TL neural net case, instead, is at least 10 times faster. On the other hand, Fig. 10(c) shows $E_{MAC\text{-}eff}$ as a function of the transmitter and receiver operation frequency. The energy consumption can be decreased 5 pJ for all the cases, when increasing the speed of the transmitter and receiver from 10 to 100 GHz, due to the reduction of the total computing time. Improvements of the SOA performance are though needed to enable high-speed all-optical signal processing: This is considered possible when exploiting concepts such as quantum dot SOAs[55] or SOAs with the carrier reservoir layer,[56] for which carrier recovery times down to 0.5–10 ps have been demonstrated, which can facilitate operation bandwidth up to 100 GHz.

Finally, we tune the power of the biased weighting elements to see the energy consumption for 64-input 64-neuron ten-layer implementation with a transceiver speed of 10 GHz. Figure 10(d) illustrates the resulting $E_{MAC\text{-}eff}$ when changing the power of the weighting elements from 0 to 30 mW (solid lines). The energy consumption per MAC rises linearly with the weight power from 8 to 14 pJ for the single layer cases and from 6 to 12 pJ for the AO-TL DNN so that the use of an all-optical multi-layer network gives a 14% improvement in effective energy consumption per MAC, with respect to E/O/E and the AO-1L implementations. In addition, the dashed lines show $E_{MAC\text{-}eff}$ for the case when a non-volatile weighting element is used, such as phase change materials:[23] For single-layer cases, the power consumption is 2.4 pJ/MAC, while for the AO-TL, an energy consumption as low as 0.7 pJ/MAC is calculated. This energy is non-zero because of the transceiver and the post-processing on the FPGA, as shown in Eqs. (1)–(3) (setting $P_w = 0$ and $P_{DAC} = 0$). This result suggests that the current control of the weighting elements contributes 5.3 pJ/MAC more for all the cases and that the SOA weighting consumes 6.3 pJ/MAC (obtained subtracting the energy consumption at 0 mW from the energy consumption at 30 mW). When substituting volatile and current biased elements with non-volatile elements in the AO-10L neural network, we can reach up to 94% energy saving for each MAC operation. In any case, the energy consumption for AO-TL neural network outperforms single-layer neural network system implementations.

VI. CONCLUSION

We analyze the performance of an all-optical neural network structure with WDM connectivity and SOA-based all-optical neurons. The linear neural network can be easily scaled as a function of WDM signals for multi-synapsis neurons: the linear processing unit can scale up to 64 c while guaranteeing a large input dynamic range under neglectable error introduction. A fully monolithically integrated all-optical neuron is experimentally demonstrated exploiting an SOA WC-based optical nonlinear function based on cross-gain modulation. The performance of the fully integrated all-optical neuron is 10% better than the hybrid case in terms of error introduction. The all-optical neural network is simulated with noise induction for benchmarking the inference of a noisy DNN built for the MNIST handwritten digit classification problem, showing that, working with 10 GS/s inputs, the all-optical approach is about 2.5 times faster than the state-of-the-art electronic GPU while guaranteeing similar accuracies.

Furthermore, we emulate the complete end-to-end system by introducing in the overall system performance calculation also the contribution of a control unit, transmitter and receiver units, together with D/A and A/D converters. The energy consumption is analyzed at a system level when an N-input M-neuron T-layers DNN is implemented. The calculation results show that the effective energy per MAC operation for an all-optical connected DNN always outperforms the single-layer DNN system. Eventually, the energy efficiency results are constrained by the speed and power consumption of the electronic side, including the DAC/ADC at the transceivers and the control FPGA for the pattern generation and signal processing, when we increase the number of synapses/neuron. Nevertheless, the AONN still performs more than 2 times better than state-of-the-art GPUs at the server level, excluding the energy for the cooling.

ACKNOWLEDGMENTS

This research was supported by the Netherlands Organization of Scientific Research (NWO) under the Zwaartekracht programma, “Research Centre for Integrated Nanophotonics”, Grant No. 024.002.033. The authors thank K. Prifti and E. Magalhães for their support on the experiment of the photonic integrated wavelength converter. The authors would also thank B. Pan for the discussion on FPGA operation.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] J. D. Kendall and S. Kumar, "The building blocks of a brain-inspired computer," Appl. Phys. Rev. **7**, 011305 (2020).

[2] F. Akopyan *et al.*, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **34**(10), 1537 (2015).

[3] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker project," Proc. IEEE **102**(5), 652–665 (2014).

[4] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," IEEE Micro **38**(1), 82 (2018).

[5] B. V. Benjamin *et al.*, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," Proc. IEEE **102**(5), 699–716 (2014).

[6] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proceedings of the International Symposium Circuits Systems* (IEEE, 2010), pp. 1947–1950.

[7] S. Han *et al.*, "EIE: Efficient inference engine on compressed deep neural network," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* (IEEE, 2016), pp. 243–254.

[8] J. Pei *et al.*, "Towards artificial general intelligence with hybrid Tianjic chip architecture," Nature **572**(7767), 106–111 (2019).

[9] A. Tavanaei *et al.*, "Deep learning in spiking neural networks," Neural Networks **111**, 47–63 (2019).

[10] U. Markowska-Kaczmar and M. Koldowski, "Spiking neural network vs multilayer perceptron: Who is the winner in the racing car computer game," Soft Comput. **19**(12), 3465–3478 (2015).

[11] See https://www.mythic-ai.com/technology/ for Mythic's chip architecture.

[12] Y. Shen *et al.*, "Silicon photonics for extreme scale systems," J. Lightwave Technol. **37**(2), 245–259 (2019).

[13] K.-I. Kitayama, M. Notomi, M. Naruse, K. Inoue, S. Kawakami, and A. Uchida, "Novel Frontier of photonics for data processing—Photonic accelerator," APL Photonics **4**, 090901 (2019).

[14] R. Stabile, G. Dabos, C. Vagionas, B. Shi, N. Calabretta, and N. Pleros, "Neuromorphic photonics: 2D or not 2D?," J. Appl. Phys. **129**(20), 200901 (2021).

[15] B. J. Shastri *et al.*, "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics **15**(2), 102–114 (2021).

[16] P. Dong, W. Qian, H. Liang, R. Shafiiha, N.-N. Feng, D. Feng, X. Zheng, A. V. Krishnamoorthy, and M. Asghari, "Low power and compact reconfigurable multiplexing devices based on silicon microring resonators," Opt. Express **18**, 9852–9858 (2010).

[17] P. P. Absil, P. Verheyen, P. De Heyn, M. Pantouvaki, G. Lepage, J. De Coster, and J. Van Campenhout, "Silicon photonics integrated circuits: A manufacturing platform for high density, low power optical I/O's," Opt. Express **23**, 9369–9378 (2015).

[18] R. Stabile, A. Rohit, and K. A. Williams, "Monolithically integrated 8 × 8 space and wavelength selective cross-connect," J. Lightwave Technol. **32**(2), 201–207 (2014).

[19] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**(7), 441–446 (2017).

[20] G. Mourgias-Alexandris, A. Totović, A. Tsakyridis, N. Passalis, K. Vyrsokinos, A. Tefas, and N. Pleros, "Neuromorphic photonics with coherent linear neurons using dual-IQ modulation cells," J. Lightwave Technol. **38**(4), 811–819 (2019).

[21] 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. Lightwave Technol. **32**(21), 4029 (2014).

[22] K. Vandoorne *et al.*, "Experimental demonstration of reservoir computing on a silicon photonics chip," Nat. Commun. **5**, 3541 (2014).

[23] J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**(7840), 52–58 (2021).

[24] R. Stabile, N. Calabretta, and B. Shi, "Large-scale photonic integrated cross-connects for optical communication and computation," in *Optical Fiber Communication Conference* (OSA, 2020), p. Th3B.1.

[25] B. Shi, N. Calabretta, and R. Stabile, "Deep neural network through an InP SOA-based photonic integrated cross-connect," IEEE J. Sel. Top. Quantum Electron. **26**(1), 7701111 (2020).

[26] B. J. Shastri *et al.*, "Principles of neuromorphic photonics," in *Unconventional Computing* (Springer, New York, 2018), pp. 83–118.

[27] B. Shi *et al.*, "SOA-based photonic integrated deep neural networks for image classification," in *Conference on Lasers and Electro-Optics* (OSA, 2019), p. SF1N.5.

[28] G. Mourgias-Alexandris *et al.*, "An all-optical neuron with sigmoid activation function," Opt. Express **27**(7), 9620 (2019).

[29] B. Shi, N. Calabretta, and R. Stabile, "Numerical simulation of an InP photonic integrated cross-connect for deep neural networks on chip," Appl. Sci. **10**(2), 474 (2020).

[30] M. Usami, M. Tsurusawa, and Y. Matsushima, "Mechanism for reducing recovery time of optical nonlinearity in semiconductor laser amplifier," Appl. Phys. Lett. **72**(21), 2657 (1998).

[31] A. Mecozzi, S. Scotti, A. D'Ottavi, E. Iannone, and P. Spano, "Four-wave mixing in traveling-wave semiconductor amplifiers," IEEE J. Quantum Electron. **31**(4), 689–699 (1995).

[32] M. J. Connelly, *Semiconductor Optical Amplifiers* (Kluwer Academic Publishers, Boston, 2002).

[33] A. Bogoni and L. Poti, "Effective channel allocation to reduce inband FWM crosstalk in DWDM transmission systems," IEEE J. Sel. Top. Quantum Electron. **10**(2), 387–392 (2004).

[34] Y. Guo, B. Aazhang, and J. F. Young, "Wavelength encoding to reduce four-wave mixing crosstalk in multi-wavelength channels," in *Conference Proceedings–Lasers Electro-Optics Society Annual Meeting* (IEEE, 1997), Vol. 2, pp. 230–231.

[35] Q. Cheng, R. Stabile, A. Rohit, A. Wonfor, R. V. Penty, I. H. White, and K. A. Williams, "First demonstration of automated control and assessment of a dynamically reconfigured monolithic 8 × 8 wavelength-and-space switch," J. Opt. Commun. Networking **7**(3), A388–A395 (2015).

[36] D. D'Agostino, D. Lenstra, H. P. M. M. Ambrosius, and M. K. Smit, "Widely tunable coupled cavity laser based on a Michelson interferometer with doubled free spectral range," in *Optical Fiber Communication Conference* (OSA, 2015), p. M2D.4.

[37] B. Shi, K. Prifti, E. Magalhães, N. Calabretta, and R. Stabile, "Lossless monolithically integrated photonic InP neuron for all-optical computation," in *Optical Fiber Communication Conference (OFC)* (OSA, 2020), p. W2A.12.

[38] B. Shi, N. Calabretta, and R. Stabile, "First demonstration of a two-layer all-optical neural network by using photonic integrated chips and SOAs," in *45th European Conference on Optical Communication (ECOC 2019)* (IET, 2019), p. 398.

[39] B. Shi, N. Calabretta, and R. Stabile, "Multi-wavelength, multi-level inputs for an all-optical SOA-based neuron," in *Conference on Lasers and Electro-Optics (CLEO)* (OSA, 2021), p. SM1B.4.

[40] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)* (USENIX association, 2016), pp. 265–283.

[41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE **86**(11), 2278–2323 (1998).

[42] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR 2015–Conference Track Proceedings)* (ICLR, 2015), pp. 1–15.

[43] A. M. de Melo and K. Petermann, "On the amplified spontaneous emission noise modeling of semiconductor optical amplifiers," Opt. Commun. **281**(18), 4598–4605 (2008).

[44] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "NVIDIA A100 tensor core GPU: Performance and innovation," IEEE Micro **41**(2), 29–35 (2021).

[45] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings–International Symposium on Computer Architecture* (ACM, 2017), Vol. Part F1286, pp. 1–12.

[46] H. Isono, "Latest standardization trend for high-speed optical transceivers with a view of beyond tera era," Proc. SPIE **11308**, 1130808 (2020).

[47] K. Hosseini *et al.*, "8 tbps co-packaged FPGA and silicon photonics optical IO," in *Optical Fiber Communication Conference 2021* (OSA, 2021), p. Th4A.2.

[48] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, "Optimizing FPGA-based accelerator design for deep convolutional neural networks," in *ACM/SIGDA International Symposium on FPGA* (ACM, 2015), pp. 161–170.

[49] M. Erett *et al.*, "A 0.5-16.3 Gbps multi-standard serial transceiver with 219 mW/channel in 16 nm FinFET," in *European Solid-State Circuits Conference* (IEEE, 2016), pp. 297–300.

[50] A. Amara, F. Amiel, and T. Ea, "FPGA vs. ASIC for low power applications," Microelectron. J. **37**(8), 669–677 (2006).

[51] S. Tanaka, S.-H. Jeong, S. Sekiguchi, T. Kurahashi, Y. Tanaka, and K. Morito, "High-output-power, single-wavelength silicon hybrid laser using precise flip-chip bonding technology," Opt. Express **20**(27), 28057 (2012).

[52] X. Zheng *et al.*, "A high-speed, tunable silicon photonic ring modulator integrated with ultra-efficient active wavelength control," Opt. Express **22**(10), 12628 (2014).

[53] E. Swindlehurst *et al.*, "An 8-bit 10-GHz 21-mW time-interleaved SAR ADC with grouped DAC capacitors and dual-path bootstrapped switch," IEEE Solid-State Circuits Lett. **2**(9), 83–86 (2019).

[54] F. Y. Liu *et al.*, "10-Gbps, 5.3-mW optical transmitter and receiver circuits in 40-nm CMOS," IEEE J. Solid-State Circuits **47**(9), 2049–2067 (2012).

[55] M. Sugawara, T. Akiyama, N. Hatori, Y. Nakata, H. Ebe, and H. Ishikawa, "Quantum-dot semiconductor optical amplifiers for high-bit-rate signal processing up to 160 Gb s^{-1} and a new scheme of 3R regenerators," Meas. Sci. Technol. **13**(11), 1683–1691 (2002).

[56] H. Sun, Q. Wang, H. Dong, G. Zhu, N. K. Dutta, and J. Jaques, "Gain dynamics and saturation property of a semiconductor optical amplifier with a carrier reservoir," IEEE Photonics Technol. Lett. **18**(1), 196–198 (2006).