Neuromorphic Silicon Photonics and Hardware-aware Deep Learning for High-Speed Inference

Miltiadis Moralis-Pegios, George Mourgias-Alexandris, Apostolos Tsakyridis, George Giamougiannis, Angelina Totovic, George Dabos, Nikolaos Passalis, Manos Kirtas, T. Rutirawut, F. Y. Gardes, Anastasios Tefas and Nikos Pleros

The relentless growth of Artificial Intelligence (AI) workloads has fueled the drive towards non-Von Neuman architectures and custom computing hardware. Neuromorphic photonic engines aspire to synergize the low-power and high-bandwidth credentials of light-based deployments with novel architectures, towards surpassing the computing performance of their electronic counterparts. In this paper, we review recent progress in integrated photonic neuromorphic architectures and analyze the architectural and photonic hardware-based factors that limit their performance. Subsequently, we present our approach towards transforming silicon coherent neuromorphic layouts into high-speed and high-accuracy Deep Learning (DL) engines by combining robust architectures with hardware-aware DL training. Circuit robustness is ensured through a crossbar layout that circumvents insertion loss and fidelity constraints of state-of-the-art linear optical designs. Concurrently, we employ DL training models adapted to the underlying photonic hardware, incorporating noise- and bandwidth-limitations together with the supported activation function directly into Neural Network (NN) training. We validate experimentally the high-speed and high-accuracy advantages of hardware-aware DL models when combined with robust architectures through a SiPho prototype implementing a single column of a 4:4 photonic crossbar. This was utilized as the pen-ultimate hidden layer of a NN, revealing up to 5.93% accuracy improvement at 5GMAC/sec/axon when noise-aware training is enforced and allowing accuracies of 99.15% and 79.8% for the MNIST and CIFAR-10 classification tasks. Channel-aware training was then demonstrated by integrating the frequency response of the photonic hardware in NN training, with its experimental validation with the MNIST dataset revealing an accuracy increase of 12.93% at a record-high rate of 25GMAC/sec/axon.

***Index Terms*—neural networks, neuromorphic computing, neuromorphic photonics, optical neural network accelerators.**

I. Introduction

THE sensational success of DL-based [1] NNs in tackling a wide range of problems, has revolutionized a wide range of applications, including e.g. image processing, computer vision [2] and bioinformatics [3], and has rekindled the scientific community's interest in AI. This success, however, was not solely based on theoretical advances, yet leveraged the availability of huge training data sets [4] and the constant growth of computing power. With Moore's Law apparent slow-down during the last decade [5], custom hardware and brain-inspired computing architectures are expected to be the upcoming major drivers of processing power growth during the next decade [6].

In this context, the photonic research community has been heavily investigating optical-based neuromorphic hardware [7]-[21], aspiring to capitalize on the same set of properties that fueled the development and eventual dominance of optical interconnects in the network domain i.e. (i) the high-bandwidth of optical components [22], that can in principle allow for tens of giga Multiply-Accumulate (MAC) operations per second (ii) the broadband signal carrying capabilities of light, that can be multiplexed in mode, polarization and wavelength [13] and act effectively as a processing power multiplication factor (iii) the low-energy and high footprint efficiency photonic components that can allow for ultra-dense, low-power deployments [23]-[25]. These advantages have shaped the neuromorphic photonics roadmaps, heralding orders of magnitude improvements compared to the electronic implementations with computational energy and area efficiency estimations predicted to reach a few fJ/MAC and >TMAC/sec/mm^2, respectively [26]-[28]. Turning these expectations into a tangible reality requires, however, a synergistic co-design and co-development roadmap among all constituent scientific and technological fields, extending from the underlying linear optical theory and architectures to the specific component design, ensuring at all times seamless integration of the photonic hardware idiosyncrasy in the DL methods and designs.

In this paper, we review recent progress in integrated photonic neuromorphic architectures and associated challenges, discussing both architectural and hardware-based performance degradation factors. We discuss the different origins of error in

Manuscript received XXXXXXX XX, XX; revised XXXX XX, XX; accepted XXXX XX, XX. This work was supported by the EC through H2020 Projects PLASMONIAC (871391) and by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of highcost research equipment grant" (Project number 4233, project acronym DeepLight).

Miltiadis Moralis-Pegios, Angelina Totovic, George Dabos, Apostolos Tsakyridis, George Giamougiannis, George Mourgias-Alexandris, Nikolaos Passalis, Manos Kirtas, Anastasios Tefas and Nikos Pleros are with the Dept. of Informatics and Center for Interdisciplinary Research and Innovation, Aristotle University of Thessaloniki, 57001, Greece (e-mail: mmoralis@csd.auth.gr, angelina@auth.gr, ntamposg@csd.auth.gr, atsakyrid@csd.auth.gr, giamouge@csd.auth.gr mourgias@csd.auth.gr, passalis@csd.auth.gr, eakirtas@csd.auth.gr, tefas@csd.auth.gr, npleros@csd.auth.gr).

T. Rutirawut and F. Gardes are with the Optoelectronics Research Centre, University of Southampton, Southampton, SO17 1BJ, UK.

experimentally realized matrix-vector multiplication (MVM) engines and we present our approach for counteracting physical errors towards high-performance coherent silicon photonic DL engines, combining robust circuit architectures together with DL training models that are optimally adapted to the idiosyncrasy of the underlying photonic hardware. On the architectural level, robustness is offered by a novel universal optical linear crossbar layout [29] that has been inspired by relevant configurations adopted by analog electronic in-memory computing and recent advances in optical MVM theory [30]. We validate its high-performance credentials through a SiPho integrated Photonic Neural Network (PNN) prototype that realizes the first column of a 4:4 crossbar (Xbar), benchmarking its performance in the well-known MNIST [19] and CIFAR-10 classification tasks at 5 and 10 GMAC/sec/axon compute line-rates, i.e., >6 orders of magnitude higher than respective state-of-the-art coherent NN layouts. This SiPho layout is then combined with hardware-aware training models where a $\sin^2(x)$ photonically implemented non-linear activation function together with noise- and bandwidth-induced limitations are a priori included in the training process, allowing in this way for additional accuracy and compute rate per axon gains. Noise-aware training is experimentally shown to allow for up to 5.93% accuracy improvement at 5 GMAC/sec/axon, yielding accuracy values of up to 99.15% and 79.8% for the MNIST and CIFAR-10 classification tasks, respectively. Additional noise originating from bandwidth limiting effects can be counteracted through a channel-response-aware training model. In this case, the frequency response of the photonic NN hardware gets embedded into the DL training process and an experimentally obtained accuracy increase of 12.93% at a record-high compute rate of 25 GMAC/sec/axon is presented for the MNIST dataset [20]. This underlines a photonic NN compute rate per axon that is >3.5x higher than the available 3-dB bandwidth of 7 GHz supported by the SiPho neuron prototype.

II. State-of-the-Art Overview and Performance Degradation Factors in Neuromorphic Photonics

The roadmap of neuromorphic photonics towards meeting the high computational power and computational area efficiency expectations has to proceed along integrated photonic solutions and high computational rates per axon,_i.e., per each neuron synapse, maintaining at the same time bit precision of at least 5-bits [26]-[28]. At the same time, their architectural layout has to support rather large NxN weight matrix deployments, so that the data signals remain as much as possible in the optical domain in order to provide the highest possible number of MAC operations for a given energy consumed by the photonic fan-in and photonic reception site. This can be easily identified by assuming an optical NxN weight matrix that gets multiplied by an N:1 optical input vector and yields an N:1 optical output vector, where every weighing node consumes a power of P_W Watts and an area of A_W mm^2, every input optical modulator used for producing the optical input vector consumes a power of P_X Watts and an area of A_X mm^2, and every photonic receiver circuit consumes P_Y Watts and has a footprint of A_Y mm^2. In such a scenario, the total energy consumed equals $NP_X + NP_Y + N^2P_W$ Watts. Assuming an operation at B MAC/sec compute rate per axon, then the total compute rate equals N^2B MAC/sec, leading to an energy efficiency of $(NP_X + NP_Y + N^2P_W)/(N^2B)$ J/MAC, which equals to $[(P_X + P_Y)/(NB)] + P_W/B$ and verifies that energy efficiency improves as N and B increase. Following a similar analysis for the computational area efficiency, the total area consumed by this circuit equals $NA_X + NA_Y + N^2A_W$ mm^2, suggesting an area efficiency of $(N^2B)/(NA_X + NA_Y + N^2A_W) = [NB/(A_X + A_Y)] + B/A_W$ in MAC/sec/mm^2, implying again a higher efficiency for increasing N and B. Finally, it should be noted, that the achieving higher compute rate is directly correlated to the achievable signal-to-noise ratio, as increasing the signal's bandwidth leads to an increase in noise bandwidth.

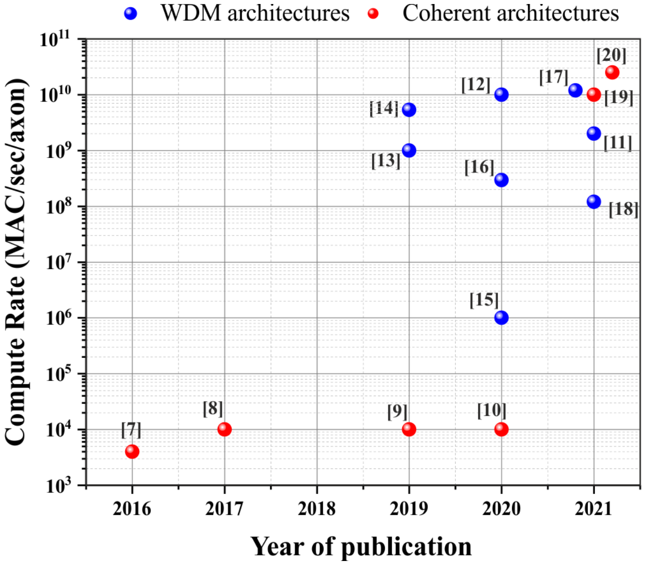

Fig. 1. Compute rate per axon performance of WDM and coherent neuromorphic architectures demonstrated experimentally between 2016-2021.

Increasing the compute rate per axon B and the circuit dimension N form actually the main approach towards competing with state-of-the-art electronic neuromorphic hardware, which utilize their high integration densities for increasing computational power through a large amount of neurons integrated on the same chip. However, a closer look into state-of-the-art neuromorphic photonic experimental deployments reveals easily that even the traditional stronghold of photonic technologies in sustaining high operational rates is not easily reflected in photonic NN layouts. Figure 1 illustrates the compute rate performance values in MAC/sec/axon reported by the rich variety of optical neural network experimental demonstrations presented within the last five years [7]-[18], with the level of integration ranging from MVM engines [10],[13],[16],[18] to fully-integrated neuron prototypes [7]-[9],[11],[12],[14],[15],[17] including both the algebraic operations and the non-linear activation function. Although different architectural schemes and different constituent integrated photonic technologies have been utilized in all these demonstrations, it can be easily observed that incoherent or Wavelength Division Multiplexing (WDM) architectures [11]-[18], were almost constantly within the GHz clock frequency operational area, allowing for a maximum of up to 11 GMAC/sec/axon compute rate accomplished only when off-chip data modulation was employed [12],[14],[17]. However, incoherent layouts typically require a different wavelength per single axon within a neuron, necessitating a high amount of wavelength resources for increasing fan-in and total computational power [13]. Single-channel optical neural networks can be accomplished only through coherent photonic interferometric layouts. This field has been until recently dominated by unitary optical linear matrix designs, where, however, the need for multiple cascaded stages of 2x2 Mach-Zehnder interferometric (MZI) meshes enforces a tight control over individual device loss uniformity and phase adjustment. This control requirement, along with the related total insertion losses of cascaded MZI stages, that becomes more pronounced when using high-speed optical modulators in the GHz range, led previous demonstration in using low speed optical fan-in and weighting technologies and therefore achieve compute rates in the sub-MHz regime [7]-[10]. Computational rates higher than 10 GMAC/sec/axon with coherent neuromorphic photonic layouts have been only recently accomplished [19]-[21], as shown in the top-right corner of Fig. 1, utilizing concepts and technologies that are overviewed in the following sections. Finally, it is worth mentioning that this comparative analysis doesn't take into account limited lab-equipment availability, comparing only the reported compute rates.

In order to identify and understand, the architectural and physical mechanisms that limit compute rate per axon performance, an analysis of both the absolute accuracy degrading constituents and the different noise sources impacting the neuromorphic photonic hardware is shown below. Fig. 2 (a) illustrates a generic layout of a WDM-based neuromorphic architecture, comprising a 4-channel input data vector, that is imprinted in the optical domain through optical modulators. The weighting functionality is typically implemented through the controlled attenuation enforced through wavelength-specific photonic components e.g., optical filters, while the weighted inputs are multiplexed through an optical multiplexer (MUX) and have finally their optical power levels summed in a photodiode. In case the targeted matrix-vector product at each output is $Y_{out,j} = \sum_0^n x_i * w_{j,i}$, possible hardware imperfections may enforce a δ_X and δ_W variation at the input data and weight values, respectively, yielding a final product of that equals:

$$Y_{Out,1} = \sum_{i=0}^{n} (x_i + \delta X_i) * (w_i + \delta_{w_i}) \qquad (1)$$

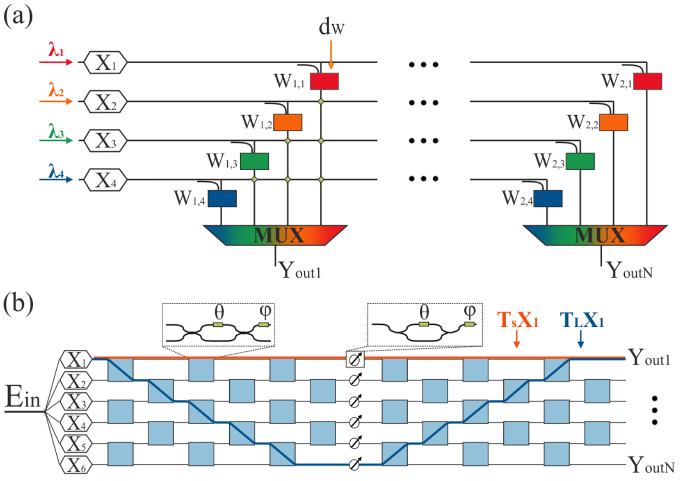

Fig. 2. (a) Generic layout of a WDM-based neuromorphic layout (b) Generic layout of a coherent-based neuromorphic layout based on the Clements rectangular mesh.

In the case of incoherent layouts, however, where every input and weight value x_i and w_i, respectively, are controlled through corresponding individual photonic modulator and weighting modules, these deviations can be compensated on a per module case by properly modifying the applied electrical driving signal. This simply implies that when input and weight values of x_i and w_i are targeted, the electrical driving voltages applied should correspond to slightly different values of $\tilde{x}_i$ and $\widetilde{w}_i$, so that $\tilde{x}_i + \delta\tilde{x}_i = x_i$ and $\widetilde{w}_i + \delta\widetilde{w}_i = w_i$, allowing for the correct values to be imprinted in the optical domain.

On the other hand, coherent-based layouts typically exploit two consecutive triangular [31] or rectangular MZI mesh layouts [32] within a Singular Value Decomposition (SVD) optical scheme, with an illustrative example of a 6-input matrix-vector multiplicator that adopts the rectangular mesh architecture depicted in Fig. 2 (b). In this case, the output signals are formed through the coherent interference of the decomposed optical beams that propagate through the cascaded MZI-stages. As such, the expected optical signal in the different outputs can be approximated for the 6-input case by $Y_{out,j} = \sum t_{i,j} X_i E_i$, with $t_{i,j}$ representing the electric field transmittivity between input i and output j, or equivalently the weight for the optical input X_i when emerging at output j. However, this architecture allows for multiple paths that connect input i and output j, suggesting that $t_{i,j}$ is formed by the sum of different transmittivity values $T^r_{i,j}$, each value corresponding to a different optical route r. This can yield an additional error factor in the experimentally realized matrix-vector product when lossy optical nodes are assumed. Considering, for example, just the extreme cases of the longest and shortest optical paths between the first input and the first output with respective transmittivity values of $T^s_{1,1}$ and $T^L_{1,1}$ when loss-less MZIs are assumed and ignoring all other possible routes between input #1 and output #1, the partial sum of the two interfering optical beams that contributes to the $t_{1,1}$ weight calculation should equal $T^s_{1,1} + T^L_{1,1}$. However, in the case of lossy optical modules, the partial sum will be given by:

$$t_{1,1-partial} = T^s_{1,1} \cdot a^7 + T^L_{1,1} \cdot a^{11} = a^7\left(T^s_{1,1} + a^4 T^L_{1,1}\right) \qquad (2)$$

, with α denoting the electrical field transmission coefficient for a single MZI and the factors α^7 and α^{11} reflecting the transmission through seven and eleven MZI nodes in the shortest and longest route, respectively. Equation (2) reveals the presence of an inherent error introduced in the weight value $t_{1,1}$ when non-zero loss MZIs are employed, which originates from the differential path losses imposed on the constituent interfering optical beams. Taking this into account and considering also the hardware-induced input and weight value deviations at every MZI building block, the total signal emerging at each output within this type of coherent layouts can

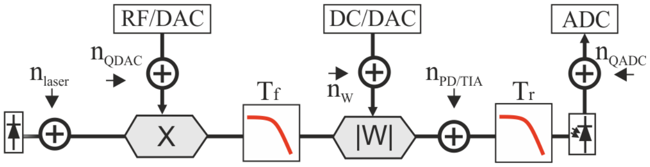

Fig. 3. Single PNN axon decomposition including main noise sources.

be written as:

$$Y_{Out,j} = \sum_{i=0}^{n}(x_i + \delta x_i) \cdot (w_i + \delta w_i + \delta_{path_{i,j}}) \qquad (3)$$

,with $\delta_{path,i,j}$ corresponding to the deviation originating from the differential path loss of the unitary coherent layout. This term can be unfortunately not fully compensated in neither the triangular nor the rectangular MZI-based unitary layouts, requiring unavoidably the use of ultra-small insertion losses for every MZI in order to minimize the error at the output signal.

In addition to the architectural and fabrication variation induced accuracy degradation, the limited frequency response and noise profiles of the underlying electro-optic hardware further degrades accuracy and operational rate of the PNN. A detailed breakdown of the different types of noise imposed on the optical signal as it traverses through a single axon is illustrated in Fig. 3. The noise that originates from the optical laser source is denoted as n_{laser} and comprises primarily the Relative Intensity Noise (RIN), producing arbitrary power level fluctuations over time [33]. The input data imprinting procedure is impacted by the quantization noise of the employed RF/DAC [34], denoted as n_{QDAC}, as well as from the bandwidth supported by the employed photonic modulator, denoted as T_f. The weight imprinting stage is considered to enforce a constant weight value when inference applications are targeted and as such can be approximated by a unity frequency response, while the quantization error of the weight imprinting DAC is denoted as n_W. At the receiver side, the TIA thermal noise and Johnson shot noise originating from the photodiode (PD) are incorporated in the $n_{PD/TIA}$ additive noise term, with the noise originating from the receiver ADC being denoted as n_{QADC}. The receiver also exhibits a typical low-pass filtering frequency response that is denoted as T_r, so that the total photonic channel response of the entire axon can be considered as having a spectral response equal to the product $T_f \bullet T_r$.

Incorporating all the above factors into our matrix-vector product analysis and taking into account that the electrical current generated at the PD output has a linear relationship with the squared of the electrical field, a generic approximation of the optical neuron output can be given by:

$$Y_{Out,j} = f\left\{\left[\sum_{i=0}^{n}[(x_i + \delta x_i) * (w_i + \delta w_i + \delta_{path_{i,j}})] \odot T_{fch}\right]^2 + \delta N\right\} \qquad (4)$$

with T_{fch} corresponding to the product of the frequency transfer functions of all constituent electronic and photonic components. The contributions of n_{QDAC} and n_W are included in the δx_i and δw_i terms, with all additional noise sources including n_{laser}, $n_{PD/TIA}$ and n_{QADC} contributing to the term δN that can be in the general case approximated by additive white gaussian noise (AWGN) [35],[36]. The function *f(·)* stands for the nonlinear activation function employed at the axon output, with a typical case for PNNs comprising $f(\cdot)=sin^2(\cdot)$ when an MZI-based modulation stage is employed at the neuron output and is driven by the ADC output signal.

Equation (4) highlights the most significant challenges that a coherent neuromorphic photonic architecture has to overcome in order to offer high accuracy performance at high operational line-rates. Minimizing inaccuracies necessitates the use of: *(i) robust circuit architectures* that can minimize the inherent noise originating from fabrication variations and the differential optical path losses, forming a loss- and fabrication-tolerant design. At the same time, this should allow for low overall insertion losses so as to enable for higher optical power levels to reach the neuron output and improved Signal-to-Noise Ratio (SNR) values, *(ii) hardware-aware DL training models*, where analog noise, bandwidth and quantization limitations together with the experimentally realized activation function can be, by default, incorporated in the training process towards building resilient models that support high-accuracy performance.

III. Robust SiPho Coherent Linear Neuron Architecture

Coherent interferometric setups relying on the unitary matrix decomposition schemes proposed by Reck [31] and Clements [32] not only lead to differential optical path losses captured by the error term of $\delta_{path_{i,j}}$ in eq.(4), but also require every weight value to depend on the control of multiple cascaded MZIs, obviously leading to the accumulation of hardware-induced deviations into an increased δw variation. We have recently proposed and demonstrated a dual-IQ-modulator-based coherent architecture [37] that allows for the direct mapping of the weight matrix values onto respective optical modules in the PNN, alleviates the need of cascaded stages of MZIs [30]-[32] and minimizes the associated layout-induced accuracy degradation. This architecture has been utilized in coherent neuromorphic photonic demonstrations and managed to break the 10 GMAC/sec/axon barrier, demonstrating computing rates of up to 25 GMAC/sec/axon as illustrated in Fig. 1. A generalized diagram of the proposed scheme for an N:1 linear neuron is illustrated in Fig. 4 (a). In this coherent photonic layout, the light entering the photonic neuron gets split in two optical beams, with the first optical beam entering a 1:N splitter to form the $X_N W_N$ dot-product and the second optical beam being forwarded to the bias branch. The bias branch is implemented through a tunable MZI followed by a Phase Shifter (PS), while the $X_N W_N$ stage comprises N identical branches, with every i-th branch imprinting the X_i data via an MZI modulator, the W_i sign through a PS, and the W_i weight absolute value through an additional tunable MZI. In the unconstrained NN implementation use case when the weight values can take any arbitrary positive or negative value, tuning of the weight sign phase-shifter and weight amplitude MZI shifter must be performed simultaneously. The $X_i W_i$ products of all optical branches recombine then again via an N:1 combiner stage to form the weighted input signal summation, which is subsequently interfering also with the bias signal to

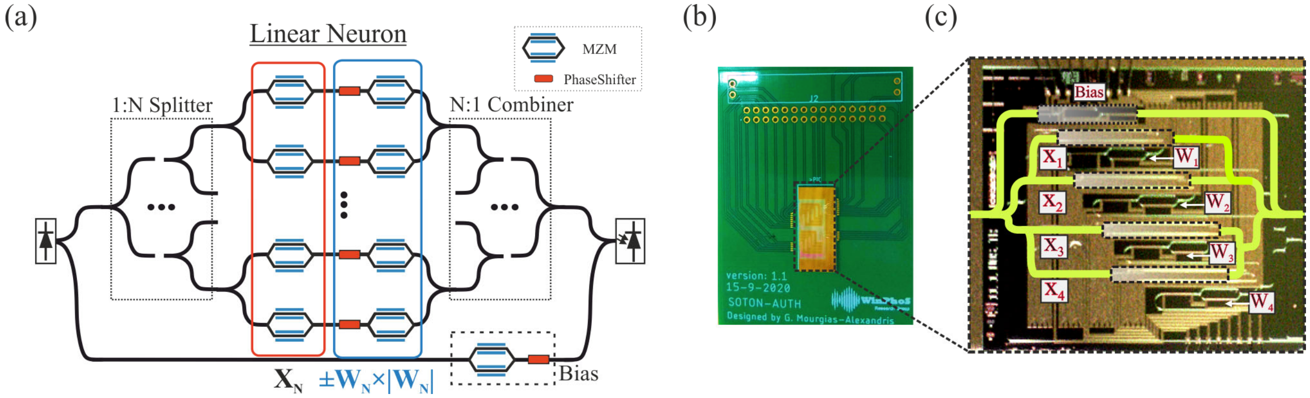

Fig. 4. (a) Schematic layout of the generic linear neuron (b) SiPho 4:1 neuron chip with up to 25 GMAC/sec/axon using EO-MZM mounted on a PCB (c) Microscope photo of the SiPho chip, with the different optical path highlighted with yellow.

produce the circuit output. The sign information of the weight value is imprinted at the phase of the light beams with φ=0 and φ=π denoting a positive and a negative value, respectively. The constructive or destructive interference between the weighted sum signal originating from the X_NW_N stage and the bias signal translates a positive sum into an optical pulse and a negative sum into a power dip around the bias signal power level.

The main advantage of our proposed architecture relies on its robustness to fabrication-related impairments that may lead to unbalanced optical losses between the different branches. Given that it allows for one-to-one mapping of the weight values into the respective optical modules at every branch, each branch can balance out both any insertion loss variation or phase mismatch through properly calibrating and fine-tuning its respective weight amplitude imprinting MZI and weight sign imprinting PS. The result of the on-chip transfer of the novel interferometric scheme is illustrated in Fig. 4 (b) and (c). Figure 4 (b) depicts the wirebonded 4:1 linear PNN mounted on a special design PCB that allows seamless access to the electrical pads of the chip. The silicon chip was fabricated in Cornerstone's Silicon Photonic 220nm platform using 1.8 mm-long asymmetric push-pull Electro-Optic (EO) MZMs for the input data imprinting stage and 500um-long Thermo-Optic (TO) PSs for the weight sign and amplitude phases. Figure 4 (c) illustrates a microscope photo of the SiPho PNN prototype, highlighting also the 4 optical paths that allow the realization of the algebraic sum: $\sum_{N=1}^{4} X_N W_N$.

Extending this single-neuron coherent interferometric layout into a multi-neuron coherent setup, where a matrix-vector multiplication $Y=W{\cdot}X$ between an N×M weight matrix W and an input N:1 vector X is realized, can be easily supported by adopting the scalability principle of the electronic Xbar layout shown in Fig. 5 (a). By splitting the optical X_i input signal into *M* identical copies and forwarding every copy into a respective N-branch weighting and N:1 recombination stage, the N:1 optical input vector *X* will get multiplied by M vectors of N-element weights or equivalently by a N×M weight matrix. This photonic Xbar layout is depicted in Fig. 5 (b) and has been theoretically validated to yield a universal linear optical operator [29],[38] that has been so far supported only through SVD-based schemes enforced over unitary optical matrix configurations [30]-[32]. It can be easily identified that this photonic Xbar allows for increased loss-tolerance compared to state-of-the-art SVD/unitary designs, given that all its M weighting columns comprise M independent weight-and-recombination stages. Given that intra-column hardware-induced deviations can be compensated through the individual control of the intra-column circuit modules, as analyzed above

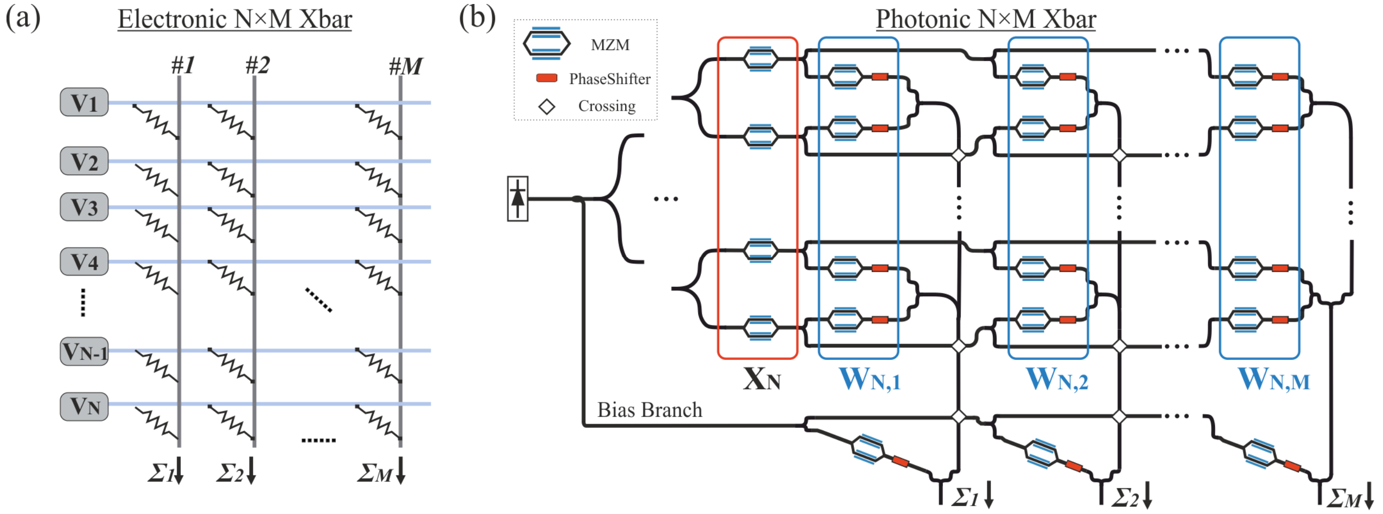

Fig. 5. (a) The electronic crossbar layout performing as the linear neural layer stage in analog electronic neural networks, (b) the corresponding analogous photonic crossbar, weight imprinting is achieved through an MZM-based attenuator while sign imprinting is achieved through a phase shifter.

for the case of the N:1 coherent neuron, the use of M independent circuit columns allows then for efficient inter-column loss-balancing by simply accommodating an M-element attenuation or amplification stage at the column outputs, concluding in this way to a fidelity-restorable design [29]. On top of that, this layout retains the main advantage of the dual-IQ-modulator-based N:1 linear neuron that requires only two cascaded MZIs for input X and weight value W imprinting. This allows for a neuron total insertion loss that scales only linearly with the weighting MZI losses [29], in direct contrast to the exponential scaling experienced by SVD-based implementations where cascaded MZIs are used in their unitary matrix constituents. The effect of this architectural-depended optical loss can be illustrated by comparing the total insertion loses of a 16×16 neuromorphic accelerator layout, following either the SVD-Clements or the Xbar layout, when the MZI losses are 1 dB. In the Clements-case the optical signal traverses in the best case and worst case: $T_s=2*\lfloor N/2\rfloor+1$, $T_l=2*N+1$ nodes respectively, resulting to a total insertion loss, excluding the input stage, ranging from 17 dB to 33 dB Considering now the Xbar layout, the optical signal traverses only a single weighting node, with the main insertion loss originating from the column/wise 1:16 splitter, that induces 12 dB of optical loss, concluding to 13 dB of total insertion loss, an improvement of 4-20 dB. This may unlock the potential for employing also high-speed photonic modulation technologies for both the input data and weight imprinting stages, which typically experience, however, higher insertion losses than the MZI weighting nodes employed so far in coherent inference engines [7]-[10]. This perspective can offer new layers of functionality for photonic neuromorphic hardware, eventually supporting also high-speed weight update rates that can be highly useful for on-chip photonic NN training engines [39]. Finally, as in both architectural approaches the total footprint is defined mainly by the optical weighting modules footprint, with their total number being N^2 for both cases, the Xbar and Clements-SVD layout share the shame footprint credentials.

The performance of the 4:1 silicon PNN illustrated in Fig. 4 (b) was experimentally benchmarked for compute rates between 1-10 GMAC/sec by replacing both neurons in the penultimate hidden layer of an NN, trained to classify a sub-section of both the handwritten digits of the MNIST dataset, as well as the tiny images of the CIFAR-10 dataset. The deployed NN model is depicted in Fig. 6 (a). It comprises 2 convolutional (CNN) layers, equipped with 32 and 64 3×3 filters respectively, followed by 3 linear layers that comprise 4, 2 and 1 linear neurons. A ReLU activation function was used in the first 3 layers, with the $\sin^2(x^2)$ being employed as the activation function in the last 2 linear layers in order to account for the photonically realized activation function when the 4:1 PNN output signal is used for driving the electro-optic MZI-based *X* input data imprinting stage of the successive PNN layer. The input samples and the respective weights of each consecutive layer denoted as x_i and w_i produce the summation $\Sigma x_i w_i$, which is in turn forwarded to the corresponding activation function *f* to realize the output $f(\Sigma x_i w_i)$. The software implementation of the proposed NN was realized in the PyTorch framework, where the network was initialized using the Xavier initialization with a gain of 2 and trained using the Adam optimizer for 20 epochs, employing a batch size of 256 samples with a learning rate of 0.0001. For the MNIST classification scenario, a subset of 11552 images, corresponding to the 3 and 5 digits, were used during the training phase, while for the CIFAR-10 case the classifier was trained to discriminate between two different tiny images. Following the training of the NN, the inference phase was performed by implementing the functionality of the 3rd Hidden Layer of the trained NN in the optical domain, as shown in the blue-highlighted section of Fig. 6 (a), where Hidden Layer#3 comprises two neurons that connect the 4 layer inputs with its 2 outputs. This was accomplished by sequentially interfacing in pairs the 4 Hidden Layer#2 outputs, as the x1, x2 and x3, x4 input signals, to the 4-channel PNN prototype and then utilizing the respective optically computed weighted sums Σ1 and Σ2 to feed the NN output layer.

Figure 6 (b) illustrates a top-down schematic diagram of the utilized photonic prototype, along with the experimental setup and deployed DSP chain blocks utilized during the evaluation procedure. A light beam at λ1=1554.55 nm was injected to the SiPho chip via a pdk-ready TE grating coupler. The $|W_{bias}|$ TO MZI was used to control the bias branch amplitude, while the PS_{bias} was used to control its phase. Two push-pull EO-MZMs were deployed to optically imprint the corresponding x_1 and x_2 values originating from the NN, while the respective $|w_1|$ and $|w_2|$ values were imprinted through TO MZIs $|w_1|$ and $|w_2|$ and their corresponding signs through calibrating the TO-PS PS_1 and PS_2, respectively. In order to interface the x1 and x2 data emerging from the NN to the integrated 4:1 PNN, their

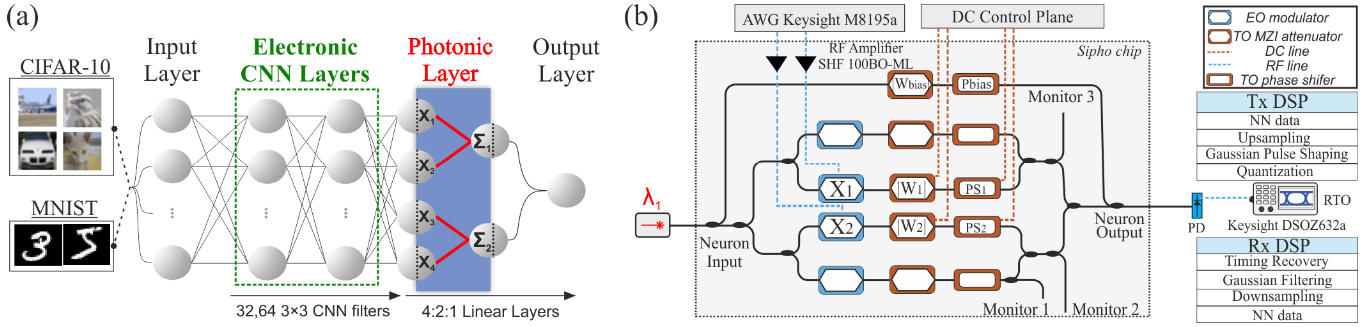

Fig. 6. (a) NN architecture for MNIST and CIFAR-10 classification, (b) Schematic diagram and experimental setup used to validate the performance of the 4-input PNN, including the deployed DSP stack

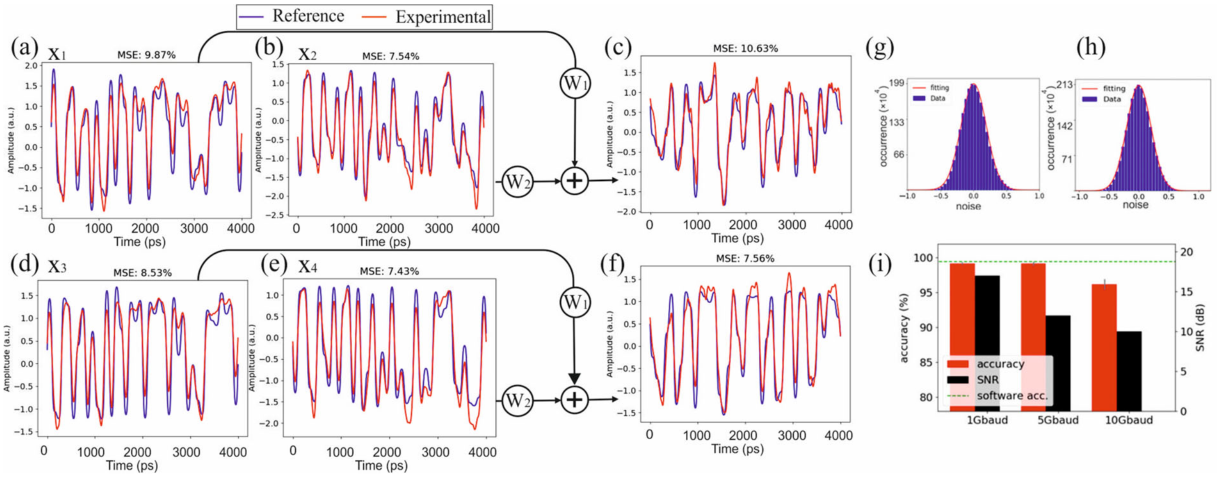

Fig. 7. Time traces of the NN's expected and experimental signals for the MNIST classification task at 5 GMAC/s/axon (a) x1, (b) x2, (c) Σ1 (d) x3, (e) x4 and (f) Σ2. (g) and (h) noise distributions of Σ1 and Σ2 (i) Accuracy and SNR measurements at 1, 5 and 10 GMAC/s/axon.

respective waveforms were upsampled from 1 to 60, 12 or 6 samples per symbol (sps), corresponding to operational data-rates of 1, 5 and 10 GSymbols/s, and were then filtered by a Gaussian filter with a band-factor of 0.8. The resulting signals were finally quantized before being uploaded to Keysight's M8195a Arbitrary Waveform Generator (AWG) operating at 60 GSa/s. The 2 output signals and their differential copies originating from the AWG, were then forwarded to 4 SHF100BO-ML RF amplifiers to drive the 2 push-pull MZMs with approximately $3V_{pp}$. The SiPho output optical signal was converted to the electrical domain by means of a PIN photodetector with 50 GHz 3dB bandwidth and was subsequently captured by a Keysight DSOZ632a Real Time Oscilloscope (RTO) with 80 GSa/s and 33 GHz bandwidth. The received signal was time-synchronized with the expected signal and was then filtered with a similar Gaussian filter before being downsampled to 1sps and forwarded to the next NN layer. The same procedure was followed for the experimental evaluation of x3 and x4 signals and their respective summation.

Figure 7 (a)-(f) illustrate the experimentally obtained results during the evaluation of the MNIST dataset, with the blue curves representing the signal originating from the NN, after being processed in the Tx DSP stack, and the red curves the experimental derived traces after Rx DSP stack processing. More specifically, Fig. 7 (a) and (b) depict the x1 and x2 signals that were interfaced to the 4:1 PNN along with the acquired experimental traces, while Fig. 7 (c) illustrates their summation x1w1+x2w2 performed in the SiPho chip, along with their expected summation result. In this case, the x1 and x2 weights were equal to w1=0.58 and w2= 0.5, while the Mean Squared Error (MSE) across the 20,987,904 samples of the experimentally obtained waveforms and their expected counterparts was 9.87%, 7.54% and 10.64% for the x1, x2 and the Σ1 signals, respectively. The same procedure was carried out for the x3 and x4 signals, with the individual traces illustrated in Fig. 7 (d) and (e) and their algebraic summation in Fig. 7 (f). In this case, the w3 and w4 weights were equal to -0.37 and 0.99, respectively, while the MSE was measured to be 8.53%, 7.43% and 7.63% for the x3, x4 and the Σ2 signals, respectively. Fig. 7 (g) and (h) illustrate the histogram of the normalized error distributions of the Σ1 and Σ2 signals in respect to the expected summation results, with both of them corresponding to a Gaussian distribution with (μ,σ) = (0, 0.20). and (μ, σ) = (0, 0.25), respectively. Finally, Fig. 7 (i) depicts the obtained mean classification accuracies on the evaluated MNIST dataset, with accuracy values of 99.2%, 99.15 % and

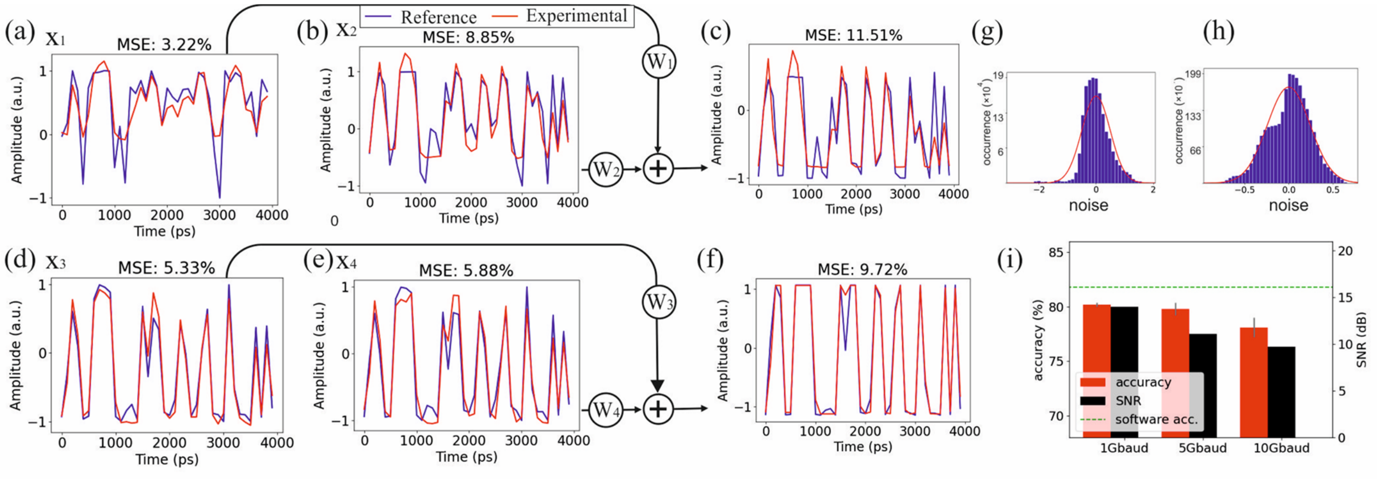

Fig. 8. Time traces of the NN's expected and experimental signals for the CIFAR classification task at 5 GMAC/s/axon (a) x1, (b) x2, (c) Σ1 (d) x3, (e) x4 and (f) Σ2. (g) and (h) noise distributions of Σ1 and Σ2 (i) Accuracy and SNR measurements at 1, 5 and 10 GMAC/sec/axon.

96.19% achieved at 1, 5 and 10 GMAC/sec/axon compute rates, respectively. The corresponding SNR values were calculated equal to 17, 12 and 10 dB, respectively.

Figure 8 (a)-(h) depict the experimental results obtained during the evaluation of the CIFAR dataset, following the same procedure as in the MNIST classification case. Fig. 8 (a) and (b) illustrate the x1 and x2 signals that were interfaced to the PNN along with the acquired experimental traces, while Fig. 8 (c) illustrates their summation x1w1+x2w2 performed in the integrated chip, along with the expected NN input signal. In this case, the x1 and x2 weights were equal to w1=0.29 and w2= 0.86, while the MSE of the experimentally derived traces and their expected counterparts was 3.22%, 8.85% and 11.51% for the x1, x2 and the Σ1 signals respectively. The same procedure was carried out for the x3 and x4 signals, with the respective results depicted in Fig. 8 (d), (e) and (f). In this case, the w3 and w4 weights were equal to -0.37 and 0.99, respectively, while the MSE was 5.33%, 5.88% and 9.72% for the x3, x4 and the Σ2 signals, respectively. Fig. 8 (g) and (h) illustrate the error distribution of the Σ1 and Σ2 signals, yielding mean and standard deviation values μ and σ that equal (μ, σ) = (0, 0.72). and (μ, σ) = (0, 0.28), respectively, when approximated by a Gaussian distribution. Finally, Fig. 8 (i) depicts the obtained classification accuracies on the CIFAR-10 dataset in comparison with the accuracy of 81.82% obtained when executed entirely by software, with accuracy values of 80.19%, 79.8% and 78.1% achieved at 1, 5 and 10 GMAC/sec/axon data rates, respectively. The corresponding SNR values were calculated equal to 14, 11.1 and 9.7 dB, respectively.

IV. Photonic Hardware-Aware Deep Learning Models

The circuit architecture of the SiPho coherent neuron allowed for up to 10 GMAC/sec/axon compute rates and for experimentally obtained accuracy values close to the reference values accomplished when executing the NN completely in the software domain, validating in this way its robustness in tolerating hardware-induced input and weight value deviations. Additional speed and accuracy improvements can be then enforced only by lifting the degradation effects of the remaining stochastic noise sources, or, alternatively, by deploying higher resilience DL training models. However, adapting DL training models and algorithms over the characteristics of the underlying analog photonic hardware has to account for a number of factors that are completely ignored in respective digital electronic NN models [40]. These include, among others, analog electro-optic noise, electro-optic bandwidth limitations, optical channel crosstalk, limited extinction ratio and value range, as well as new types of non-linear activation functions that can be realized experimentally in the electro-optical domain but typically deviate from the traditional ReLU, sigmoid etc. activation functions employed in conventional DL models. On-going research in DL models specifically tailored for PNNs has led to some first encouraging findings, having already validated the potential to train NNs with photonic $\sin^2(x^2)$ [40] and photonic sigmoid activation functions [41],[42]. On the same line, quantization noise of the constituent DAC and ADCs has been quantified and the robustness of specially trained NN models in quantization limited use cases has been also validated [43]. Moreover, the effect of non-deterministic noise sources, that were approximated as AWGN in PNNs, was also studied both in feed-forward [35],[36] as well as in Recurrent PNN layouts [44],[45] revealing that specifically trained DL models can maintain their high accuracy credentials even when deployed onto relatively noisy photonic substrates.

In this section, we present the deployment and experimental validation of DL algorithms and models tailored to the idiosyncrasy and hardware limitations of its underlying photonic platform. Two properly adapted DL models have been realized and experimentally validated over the SiPho PNN, focusing on analog electro-optic noise-aware and on channel-response-aware training. The noise-aware training takes into account the non-deterministic noise sources of the photonic hardware, targeting at noise-resilient DL models that can restore accuracy values to the level accomplished within a noise-less hardware environment. Channel-response-aware training integrates the transfer function of the electro-optic hardware within the training procedure, aiming at compute rates per axon that go significantly beyond the available electro-optic channel bandwidth.

A. Noise-aware Training

To compensate for the non-deterministic noise sources that impact signal quality in the photonic PNN, a specialized DL model that treats the various non-deterministic noise contributions as AWGN [35],[36] has been designed. Matching the training with the experimental procedure can be accomplished by introducing zero-mean AWGN during the forward propagation phase of the NN training, using a noise standard deviation value that equals the respective standard deviation of the experimentally characterized noise. Experimental characterization of the PNN noise levels is performed through the analysis of a pilot test signal that propagates through the SiPho PNN. More specifically, a 512-symbol-long pseudorandom bit sequence was transmitted through the PNN and the noise distribution of the captured output signal, comprising of 100 repetitions of the transmitted

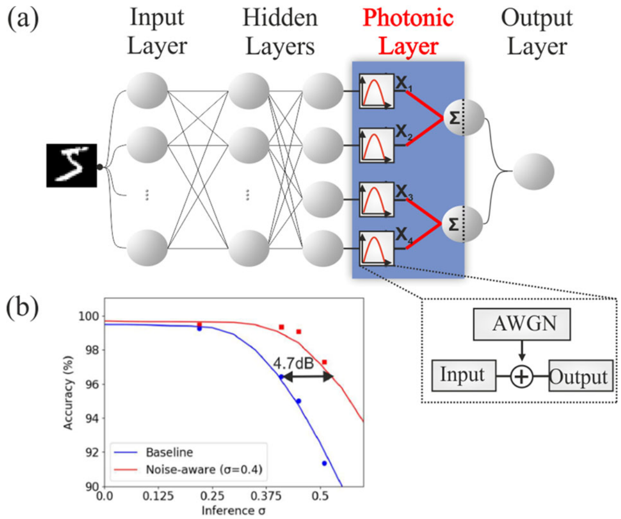

Fig. 9. (a) Modified photonic NN MNIST classifier, that incorporates AWGN sources in the training phase (b) MNIST classification accuracy versus noise standard deviation at 5 GMAC/axon/s. Solid lines represent the simulation-based results and the scatter points the experimentally derived ones.

signal, was benchmarked versus the reference electrical waveform, revealing a gaussian shaped noise distribution with a standard deviation of 0.4. The line rate of the PRBS sequence was 5 Gbit/s, while a low-pass brick wall filter with a cut-off frequency of 8 GHz was used after sampling at an RTO. The same settings were applied during the photonic implementation of the NN, to conclude to the same frequency response and noise bandwidth for both experiments. This noise distribution arises from the additive summation of the different non-deterministic noise sources and as such can approximate the non-deterministic noise behavior of the PNN. Fig. 9 (a) depicts the NN architecture that has been also used as the MNIST classifier in section 3, including the corresponding modifications in its photonic layer in order to interlace an AWGN module in its photonic axons. The retraining procedure was implemented in the PyTorch software model of the PNN, considering AWGN with a zero mean value and a standard deviation of σ=0.4 for 200 epochs.

Figure 9 (b) illustrates the comparative accuracy performance between the noise-aware and the baseline MNIST classification models for different noise levels with a standard deviation ranging between 0 and 0.6 at a compute rate of 5 GMAC/sec/axon, with the baseline model referring to the case of a noise-less environment. Increased noise levels were implemented experimentally by attenuating the power level of the neuron output signal prior reaching the receiver, resulting in this way to lower SNR values. Even though attenuating the optical power at the Rx side, attenuates also the noise originating from the transmitter side, our approach captures with high precision the noise profile of the photonic accelerator, that is dominated by laser RIN, PD shot noise and TIA thermal noise. The solid lines correspond to the accuracy levels obtained when executing MNIST classification entirely in the software domain, with the scatter points revealing the accuracy values obtained when the photonic layer is executed experimentally over the PNN. The first observation concerns the excellent matching between the experimentally obtained accuracy values depicted by the scattered rectangle points and the corresponding solid lines in both the noise-aware and baseline training models, validating the robustness of both the developed software framework and the effectiveness of the noise-aware model. The second observation regards the performance gains realized by the noise-aware model. Accuracy improvements are more pronounced as the noise level is increasing, reaching a performance gain of 5.93% when the AWGN has a standard deviation of 0.4. Performance gains are expected to be even more significant in case more NN layers are implemented in the photonic domain, since in that case the baseline model would concede to higher noise standard deviation values. It should be noted that the performance gains of the noise-aware model can be interpreted and exploited in two possible ways, opening new paths for future PNN implementations: (i) they can either translate to increased accuracy values when the baseline and the noise-aware models are executed assuming identical noise levels in both cases, or (ii) they can lead to relaxed power budget requirements for the same accuracy performance, allowing for lower PNN output optical powers and as such lower SNR values, which may form a critical advantage when amplifier-less integrated PNN solutions are employed. The latter has been also validated experimentally by increasing the noise standard deviation of the PNN circuit until the noise-aware model matches the accuracy obtained by the baseline model at the reference noise level with σ=0.4, indicating that a 4.7dB lower optical power can be utilized in this case.

B. *Channel-aware Training*

While non-deterministic noise sources can be approximated through AWGN, the Inter Symbol Interference (ISI) originating from the limited or non-linear channel response of the underlying photonic components necessitates an analytical mathematical model approximation. To this end, we developed a DL method that integrates a specially designed software building block in the NN training procedure, which allows for the inclusion of the PNN channel response in the NN training phase. Fig. 10 (a) illustrates the modified version of an NN, trained to classify the handwritten digits of the MNIST dataset, including the channel response modelling block in its photonic output layer. The data stream exiting the last hidden layer is converted to the frequency domain via Real Fast Fourier Transformation (RFFT) and gets then multiplied with an arbitrary channel response in the frequency domain. The resulting signal is then converted back to the time domain through an Inverse Real Fast Fourier Transformation (IRFFT). The PNN channel response was approximated by experimentally deriving the transfer function of the 4:1 Sipho PNN, which is illustrated in Fig. 10 (b) to follow a low-pass filtering response with a 3-dB bandwidth of approximately 7.5 GHz, mainly dictated by the response of the MZM modulator.

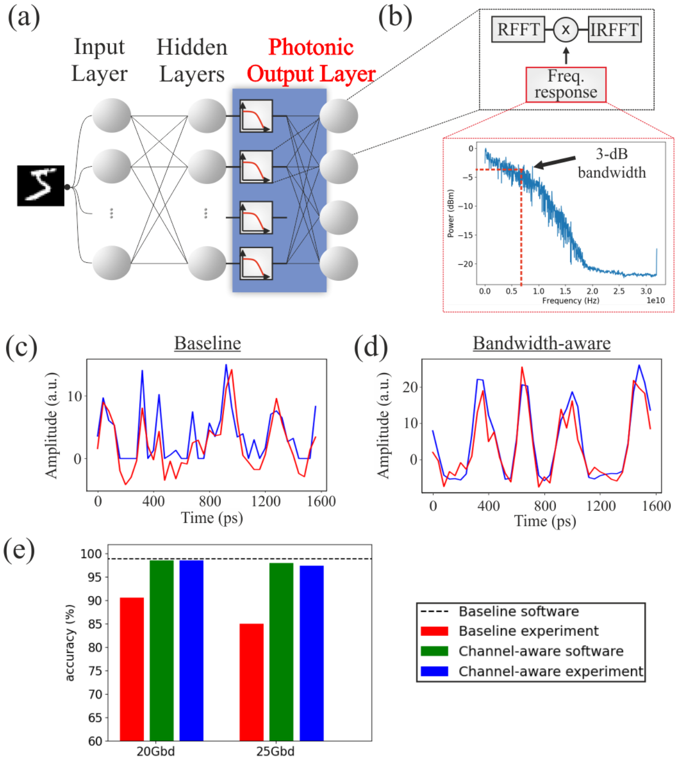

Fig. 10. (a) Modified photonic NN MNIST classifier, that incorporates the transfer function of the PNN in the training phase (b) Transfer function of the 4-input SiPHO PNN. Experimental and expected traces at 20 Gbaud for the (c) Baseline and (d) Bandwidth-aware (e) Accuracy results for both models at 20 and 25 Gbaud.

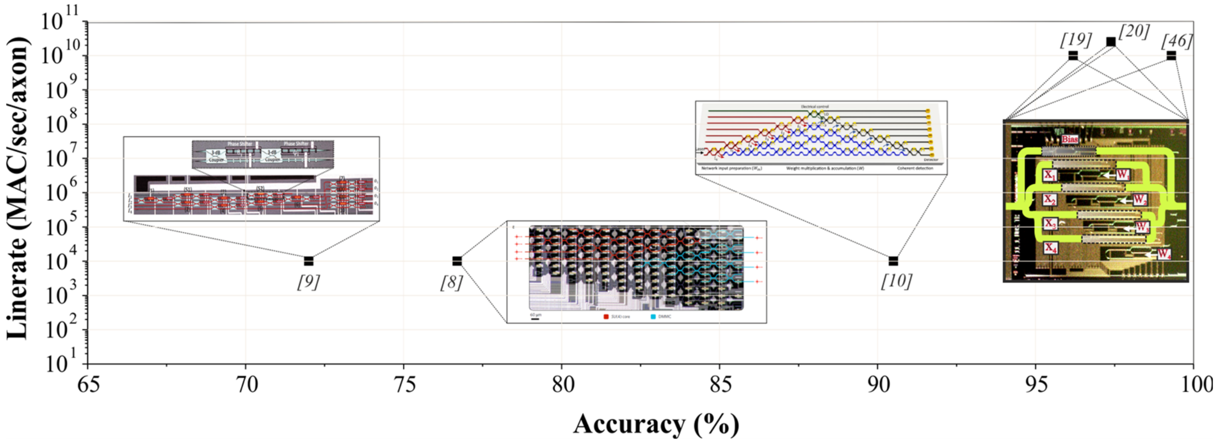

Fig. 11. Reported performance of experimentally demonstrated coherent linear neuron engines in terms of line rate and classification accuracy.

Figures 10 (c), (d) depict the experimentally obtained (red curves) versus the NN expected (blue curves) time traces for both the baseline (Fig.10 (c)) and the channel-aware (Fig.10 (d)) model at 20 Gbaud, with the baseline referring to the trained model without accounting for the channel response. It can be easily observed that the traces in the channel-aware model case follow much closer the respective expected waveform, revealing an error decrease between the received and the expected signals as we move from the baseline to the channel-aware model. This error decrease is quantitively analyzed in the reported accuracies values illustrated in Fig. 10 (e). The baseline model achieves experimental accuracies of 90.6% and 85.07% on the MNIST classification task at 20 and 25 GMAC/sec/axon, i.e., a degradation of 8.3% and 13.83% compared to the accuracy levels accomplished by the baseline model when executed entirely in the software domain. The software-derived accuracies for the channel response-aware scheme reached 98.6% and 98%, experiencing only a minor a degradation of only 0.3% and 0.9%, respectively, compared to the baseline software execution. The bandwidth-limit resilient character of the channel-response-aware model was also experimentally validated at both compute rates, managing to sustain accuracy values of 98.51% & 97.37% at 20 and 25 GMAC/sec/axon, respectively. This translates into 7.91% and 12.3% accuracy performance improvement over the corresponding experimentally executed baseline models. The experimental evaluation of the channel-response-aware model was carried out by imprinting the data input values on the SiPho PNN and recording the resulting optical waveforms, with the rest of the NN functionalities including weighting, summation and activation performed in the software domain.

V. Conclusion

We have overviewed the state-of-the-art in experimentally deployed PNN demonstrations, validating that state-of-the-art coherent PNN layouts relying on unitary optical matrix implementation schemes have still not entered the GHz operational regime. Following a detailed analysis of the different noise contributions experienced by a photonic NN hardware platform, we have reviewed a robust photonic Xbar architecture [29] that can compensate hardware-induced errors in matrix-vector multiplication tasks, while NN accuracy errors originating by stochastic noise sources should be minimized through hardware-aware DL training models. We have demonstrated a 4:1 SiPho PNN chip prototype that realizes the first column of a 4:4 Xbar architecture, validating through respective MNIST and CIFAR-10 classification experiments its credentials to drive coherent neuromorphic layouts into the 5 and 10 GHz operational line-rate regimes. In addition, we have introduced and experimentally validated the accuracy and speed performance gains enforced via noise-aware and channel-response-aware DL training models. The overall progress sustained in the field of coherent PNNs by the combined use of a robust SiPho coherent PNN architecture and hardware-aware DL models can be schematically captured in Fig. 11, which provides a pictorial representation of the MAC/sec/axon compute rate versus accuracy metrics reported by coherent PNN demonstrations so far [8]-[10]. It clearly reveals that the experimentally obtained accuracy values accomplished so far by state-of-the-art coherent PNNs ranged only between 72% and 90.5%, with the operational compute rate per axon never exceeding 10 kHz. At the same time, the proposed silicon coherent neuromorphic platform equipped with hardware-aware DL models allowed for the first time to penetrate the regime of >10GMAC/sec/axon compute rates while safeguarding >95% accuracy values, outperforming all state-of-the-art coherent neurons by ~6 orders of magnitude in terms of per axon compute rates. This may open completely new perspectives for neuromorphic photonic circuit applications, allowing coherent silicon photonic layouts to migrate to high-speed and high-accuracy inference settings that may be equipped with additional programmable or performance acceleration functions when combined with WDM capabilities [47]. This roadmap has to proceed along the lines of high-density photonic-electronic co-integration, leveraging the latest advances in high-speed driver and TIA array co-packaging, towards a functional neuromorphic accelerator prototype [48],[49]. Finally, this may even support a reliable transition into the use of PNNs for training applications when utilizing high-speed electro-optic SiPho technology at both the input and the weighting stages [21].

Acknowledgments

The authors would also like to thank Dr. Francesco Zanetto from Politecnico di Milano for wirebonding of the SiPho chip.

REFERENCES

[1] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.

[2] Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in Advances in Neural Information Processing Systems (NeurIPS, Tahoe, Nevada, 2012), pp. 1097–1105

[3] Cao, Y., Geddes, T.A., Yang, J.Y.H. et al. Ensemble deep learning in bioinformatics. Nat Mach Intell 2, 500–508 (2020).

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. 115, 211 (2015).

[5] G. E. Moore, Cramming More Components onto Integrated Circuits, Electronics 38, 114 (1965).

[6] J. D. Kendall, S. Kumar , "The building blocks of a brain-inspired computer", Applied Physics Reviews 7, 011305 (2020).

[7] A. Ribeiro et al, "Demonstration of a 4 × 4-port universal linear circuit," Optica 3, 1348-1357 (2016).

[8] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," Nat. Photonics, vol. 11, no. 7, pp. 441–446, 2017

[9] F. Shokraneh et al, "A Single Layer Neural Network Implemented by a 4x4 MZI-Based Optical Processor," in IEEE Photon. J., vol. 11, no. 6, pp. 1-12, Dec. 2019.

[10] H. Zhang et al. "An optical neural chip for implementing complex-valued neural network", Nat Commun 12, 457 (2021).

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

[12] B. Shi et al, "Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect," IEEE J. Sel.Top. Quantum Electron., vol. 26, no. 1, pp. 1–11, Jan. 2020.

[13] A. Tait et al., "Silicon Photonic Modulator Neuron", Physical Review Applied, vol. 11, no. 6, 2019.

[14] Y. Huang et al, "Programmable matrix operation with reconfigurable time-wavelength plane manipulation and dispersed time delay," Opt. Express 27, 20456-20467 (2019).

[15] C. Huang et al, "Demonstration of photonic neural network for fiber nonlinearity compensation in long-haul transmission systems," in Optical Fiber Comm. Conf. PDP 2020, Th4C.6., 2020.

[16] T. F. de Lima et al, "Real-time Operation of Silicon Photonic Neurons," in Optical Fiber Communication Conference (OFC) 2020, M2K.4., 2020.

[17] X. Xu et al., "11 TOPS photonic convolutional accelerator for optical neural networks", Nature 589, 44–51 (2021).

[18] W. Zhang et al, "Microring Weight Banks Control beyond 8.5-bits Accuracy", arXiv preprint arXiv:2104.01164, 2021

[19] G. Mourgias-Alexandris, et al., "A Silicon Photonic Coherent Neuron with 10GMAC/sec processing line-rate ", Optical Fiber Comm. Conf., Tu5H.1, 2021.

[20] G. Mourgias-Alexandris et al., "25GMAC/sec/axon photonic neural networks with 7GHz bandwidth optics through channel response-aware training," 2021 European Conference on Optical Communication (ECOC), pp. 1-4, 2021.

[21] G. Giamougiannis et al, "Silicon-integrated coherent neurons with 32GMAC/sec/axon compute line-rates using EAM-based input and weighting cells" in ECOC (2021).

[22] H. Li, G. Balamurugan, T. Kim, M. N. Sakib, R. Kumar, H. Rong, J. Jaussi, and B. Casper, "A 3-D-Integrated Silicon Photonic Microring-Based 112-Gb/s PAM-4 Transmitter With Nonlinear Equalization and Thermal Control," IEEE Journal of Solid-State Circuits 56(1), 19–29 (2021).

[23] M. Moralis-Pegios, S. Pitris, T. Alexoudi, H. Ramon, X. Yin, J. Bauwelinck, Y. Ban, P. de Heyn, J. van Campenhout, and N. Pleros, "52 km-Long Transmission Link Using a 50 Gb/s O-Band Silicon Microring Modulator Co-Packaged With a 1V-CMOS Driver," IEEE Photonics Journal 11(4), 1–7 (2019).

[24] T. Alexoudi, N. Terzenidis, S. Pitris, M. Moralis-Pegios, P. Maniotis, C. Vagionas, C. Mitsolidou, G. Mourgias-Alexandris, G.T. Kanellos, A. Miliou, K. Vyrsokinos and N. Pleros, "Optics in Computing: from Photonic Network-on-Chip to Chip-to-Chip Interconnects and Disintegrated Architectures", IEEE/OSA J. of Lightwave Technol., Vol. 37, No. 2, pp. 363-379, Jan. 2019

[25] S. Pitris, C. Mitsolidou, M. Moralis-Pegios, K. Fotiadis, Y. Ban, P. De Heyn, J. Van Campenhout, J. Lambrecht, H. Ramon, X. Yin, J. Bauwelinck, N. Pleros and T. Alexoudi, "400 Gb/s Silicon Photonic Transmitter and Routing WDM technologies for glueless 8-socket Chip-to-Chip interconnects", IEEE/OSA J. of Lightwave Technol., Vol. 38, No. 13, pp. 3366-3375, July 2020

[26] Shastri, B.J., Tait, A.N., Ferreira de Lima, T. et al. Photonics for artificial intelligence and neuromorphic computing. Nat. Photonics 15, , 2021.

[27] A. R. Totović et al, "Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap," in IEEE J. of Sel. Topics in Quantum Electronics, vol. 26, no. 5, pp. 1-15, Sept.-Oct. 2020.

[28] M. Nahmias, et. al. "Photonic Multiply-Accumulate Operations for Neural Networks," IEEE JSTQE, 26 (1), 2020.

[29] G. Giamougiannis et. al., "Coherent photonic crossbar as a universal linear operator", submitted to Laser and Photonics Review (2021).

[30] David A. B. Miller, "Self-configuring universal linear optical component [Invited]," Photon. Res. 1, 1-15 (2013).

[31] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," Phys. Rev. Lett. 73, 58–61 (1994).

[32] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. Steven Kolthammer and I. A. Walmsley, "Optimal design for universal multiport interferometers," Optica 3, 1460–1465 (2016).

[33] I. Joindot, "Measurements of relative intensity noise (RIN) in semiconductor lasers," Journal de Physique III, vol. 2, no. 9, pp. 1591–1603,1992.

[34] C. Pearson, "High-speed, analog-to-digital converter basics," Texas Instruments, Dallas, TX, USA, App. Rep. SLAA510, Jan. 2011. [Online]. Available: http://www.ti.com/lit/an/slaa510/slaa510.pdf

[35] N. Passalis, M. Kirtas, G. Mourgias-Alexandris, G. Dabos, N. Pleros, and A. Tefas, "Training noise-resilient recurrent photonic networks for financial time series analysis," Eur. Signal Process. Conf., vol. 2021-Janua, pp. 1556–1560, 2021.

[36] G. Dabos et al., "End-to-end deep learning with neuromorphic photonics," Proc. SPIE11689, 1168901 (2021).

[37] G. Mourgias-Alexandris et al., "Neuromorphic Photonics With Coherent Linear Neurons Using Dual-IQ Modulation Cells," in J. of Lightwave Technol., vol. 38, no. 4, pp. 811-819, 15 Feb.15, 2020.

[38] N. Pleros et al., "Compute with Light: Architectures, Technologies and Training Models for Neuromorphic Photonic Circuits". 2021 European Conference on Optical Communication (ECOC), pp. 1-4, 2021.

[39] H. Zhang, J. Thompson, M. Gu, X. Dong Jiang, H. Cai, P. Yang Liu, Y. Shi, Y. Zhang, M. Faeyz Karim, G. Qiang Lo, X.Luo, B.Dong, L. Chuan Kwek, and A. Qun Liu, " Efficient On-Chip Training of Optical Neural Networks Using Genetic Algorithm", ACS Photonics 2021 8 (6), 1662-1672

[40] N. Passalis, G. Mourgias-Alexandris, A. Tsakyridis, N. Pleros, A. Tefas, "Training deep photonic convolutional neural networks with sinusoidal activations," IEEE Transactions on Emerging Topics in Computational Intelligence, 2019.

[41] G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos and N. Pleros, "An all-optical neuron with sigmoid activation function", Optics Express, Vol. 27, No. 7, pp. 9620-9630, Mar. 2019

[42] J. Crnjanski, M. Krstic, A. Totovic, N. Pleros and D. Gvozdic, "Adaptive sigmoid-like and PReLU activation functions for all-optical perceptron", OSA Optics Lett., Vol. 46, No. 9, pp. 2003-2006, 2021.

[43] S. Garg, A. Jain, J. Lou, and M. Nahmias, "Confounding Tradeoffs for Neural Network Quantization," 2021, [Online]. Available: http://arxiv.org/abs/2102.06366.

[44] G. Mourgias-Alexandris, N. Passalis, G. Dabos, A. Totovic, A. Tefas and N. Pleros, "A Photonic Recurrent Neural Network for Time-Series Classification", IEEE J. of Lightwave Technol., Vol. 39, No. 5, pp. 1340-1347, Mar. 2021.

[45] G. Mourgias-Alexandris, G. Dabos, N. Passalis, A. Totović, A. Tefas and N. Pleros, "All-optical WDM Recurrent Neural Networks with Gating", IEEE J. on Sel. Topics of Quantum Electron., Vol. 26, No. 5, pp. 1-7, Sept. 2020, doi: 10.1109/JSTQE.2020.2995830.

[46] G. Mourgias-Alexandris et al., "Noise-resilient and high-speed deep learning with coherent silicon photonics", submitted at Nature Communications (2021).

[47] A. Totovic et al., "Programmable photonic neural networks through WDM-equipped coherent optics", submitted at Nature Scientific Reports (2021).

[48] M. Moralis-Pegios, S.Pitris, T. Alexoudi, et al., "A 4-channel 200 Gb/s WDM O-band Silicon Photonic Transceiver sub-assembly", Opt. Express 28, 5706-5714,2020.

[49] D. Guermandi et al., "TSV-assisted hybrid FinFET CMOS — Silicon Photonics Technology for high density optical I/O," 45th European Conference on Optical Communication (ECOC 2019), 2019.