APL Photonics

RESEARCH ARTICLE | NOVEMBER 13 2025

A 262 TOPS hyperdimensional photonic AI accelerator powered by a Si_3N_4 microcomb laser

Special Collection: Optical Computing Systems

Christos Pappas ; Antonios Prapas ; Theodoros Moschos ; Manos Kirtas ; Odysseas Asimopoulos ; Apostolos Tsakyridis ; Miltiadis Moralis-Pegios ; Chris Vagionas ; Nikolaos Passalis ; Cagri Ozdilek ; Timofey Shpakovsky; Alain Yuji Takabayashi ; John D. Jost; Maxim Karpov ; Anastasios Tefas ; Nikos Pleros

Check for updates

APL Photonics 10, 110805 (2025)
https://doi.org/10.1063/5.0271374

View Online
Export Citation

Articles You May Be Interested In

Terahertz microcomb oscillator stabilized by molecular rotation

APL Photonics (January 2024)

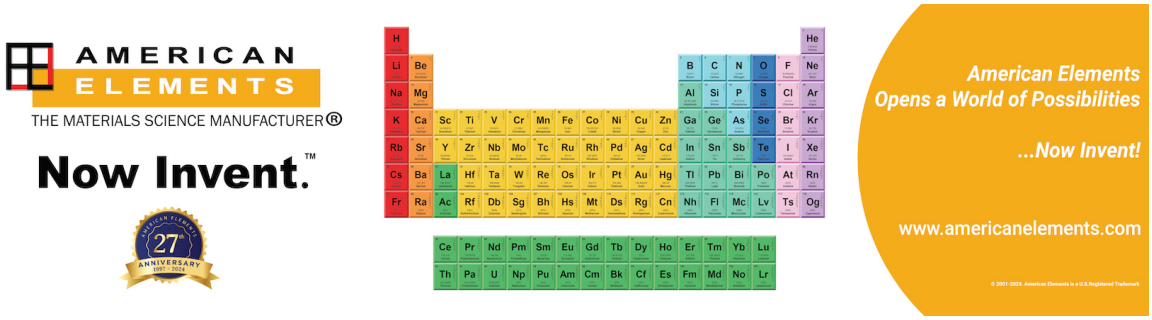
60 Gbps real-time wireless communications at 300 GHz carrier using a Kerr microcomb-based source

APL Photonics (June 2023)

Wideband image-reject RF channelization based on soliton microcombs (invited paper)

APL Photonics (September 2023)

09 December 2025 17:52:39



AIP Publishing

A 262 TOPS hyperdimensional photonic AI accelerator powered by a Si_3N_4 microcomb laser

Cite as: APL Photon. 10, 110805 (2025); doi: 10.1063/5.0271374
Submitted: 17 March 2025 • Accepted: 29 October 2025 •
Published Online: 13 November 2025

View Online Export Citation CrossMark

Christos Pappas,[1,2,a] Antonios Prapas,[1,2] Theodoros Moschos,[1,2] Manos Kirtas,[1,3] Odysseas Asimopoulos,[2,4] Apostolos Tsakyridis,[1,2] Miltiadis Moralis-Pegios,[1,2] Chris Vagionas,[1,2] Nikolaos Passalis,[3,5] Cagri Ozdilek,[6] Timofey Shpakovsky,[6] Alain Yuji Takabayashi,[6] John D. Jost,[6] Maxim Karpov,[6] Anastasios Tefas,[1,3] and Nikos Pleros[1,2]

AFFILIATIONS

[1] Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[2] Center for Interdisciplinary Research and Innovation, Balkan Center, Thermi 57001, Greece
[3] Computational Intelligence and Deep Learning Group, AUTH, 54124 Thessaloniki, Greece
[4] Department of Physics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[5] Department of Chemical Engineer, Faculty of Engineer, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[6] Enlightra, Rue de Lausanne 64, Renens 1020, VD, Switzerland

Note: This paper is part of the Special Topic on Optical Computing Systems.
[a] **Author to whom correspondence should be addressed:** chripapp@csd.auth.gr

ABSTRACT

The ever-increasing volume of data demarcating from the exponential scale of Artificial Intelligence (AI) and Deep Learning (DL) models motivated research into specialized AI accelerators in order to complement digital processors. Photonic Neural Networks (PNNs), with their unique ability to capitalize on the interplay of multiple physical dimensions, including time, wavelength, and space, have been brought forward with a credible promise for boosting computational power and energy efficiency in AI processors. In this article, we experimentally demonstrate a novel multidimensional arrayed waveguide grating router (AWGR)-based photonic AI accelerator that can offload bandwidth-bounded linear algebra while leaving memory hierarchy, control, and nonlinearities to electronics and can execute tensor multiplications at a record-high total computational power of 262 TOPS, offering a ~24× improvement over the existing waveguide-based optical accelerators. It consists of a 16 × 16 AWGR that exploits the time-, wavelength-, and space-division multiplexing (T-W-SDM) for weight and input encoding, together with an integrated Si_3N_4-based frequency comb for multi-wavelength generation. The photonic AI accelerator has been experimentally validated in both Fully Connected (FC) and Convolutional NN (CNN) models, with the FC and CNN being trained for DDoS attack identification and MNIST classification, respectively. The experimental inference at 32 Gbaud achieved a Cohen's kappa score of 0.8652 for DDoS detection and an accuracy of 92.14% for MNIST classification, respectively, closely matching the software performance.

© 2025 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/5.0271374

I. INTRODUCTION

Electronic AI accelerators continue to improve at the system level, yet the slowing of traditional voltage/frequency scaling ("power wall," "dark silicon")[1] together with the rapid rise in compute demand for frontier AI training[2,3] motivated a paradigm shift into specialized hardware that can complement digital platforms, sustaining the compute and energy growth. Within this framework, photonic neural networks (PNNs) have been theoretically predicted to hold the credentials for addressing these computational and energy challenges toward enabling peta-scale

09 December 2025 17:52:39

compute power and fJ/OP-scale energy efficiency,[4–6] leveraging their high-bandwidth and low-power characteristics with their constantly growing integration maturity. However, the transfer of theoretical predictions into experimental demonstrations[6–31] has revealed a different reality for PNNs, primarily marked by architectural challenges toward safeguarding scalable layouts. Given the footprint handicap of integrated photonic components against their electronic counterparts, PNNs can address scalability only through architectural schemes that can seamlessly support the interplay between time- and space-dimensions together with the traditional optical advantages, i.e., the wavelength-dimension and high-speed operation, which, however, is still absent in state-of-the-art integrated photonic Matrix-Vector-Multiply (MVM) architectures.

More specifically, the demonstrations in Refs. 7 and 16 exploit photonic meshes and spatial division multiplexing (SDM), executing computations via the use of cascaded Mach–Zehnder interferometer (MZI) nodes. This approach directly correlates physical with computational space, implying a scalability plateau since the larger networks would inevitably result in higher insertion losses.[32] On the other hand, extensive research has been conducted on PNN architectures that utilize the wavelength division multiplexing (WDM) technique.[15,26] The main building block in these architectures is the microring resonator (MRR) bank, which comprises multiple MRRs flanked by two parallel waveguides and is responsible for implementing channel-selective weighting. Despite their impressive performance in a range of applications, the computational power and circuit size of these layouts can scale only by increasing the number of laser sources. At the same time, they necessitate the simultaneous operation and precise control of numerous resonant devices, raising an additional concern regarding their power consumption and scalability perspectives. Photonic MVM circuit size can enjoy unlimited scaling only by exploiting time-division-multiplexing (TDM), with the authors in Refs. 12, 14, 20, and 27 demonstrating the effective synergy of SDM and TDM by incorporating high-speed input/weight nodes and time integrating receivers for the accumulation operation. Consequently, the size of the PNN is effectively increased without the need to fabricate large photonic circuits, while the employment of the time integrating receiver enables the use of low-power and low-cost analog-to-digital converters (ADCs) as it relaxes their bandwidth requirements. Yet, these architectures neglect the use of wavelength division multiplexing (WDM) that forms the typical capacity parallelization factor in optics, limiting the ability of PNNs to fully exploit all available degrees of freedom and hence potentially boost their computational power and energy efficiency metrics. The benefits of additionally incorporating the wavelength dimension in time-space division multiplexed setups have been highlighted in more recent demonstrations,[23,28,29] leading also to computational powers up to 11 TOPS[23] that comprise the current record among all state-of-the-art waveguide-based optical processors so far. To the best of our knowledge, only diffractive optics have managed to break the 100 TOPS computational power barrier,[33–35] either by chiplets built upon diffractive principles[33] or by a hybrid combination of diffractive regions with inference modules.[34,35] However, the use of diffractive elements requires the *a priori* encoding of weighting values during the fabrication process, and as such, it inherently constrains photonic AI accelerators to a certain inference task. To this end, the high computational powers enabled by diffraction-based optical processors come at the expense of their flexibility and universality, restricting their deployment only in application-specific inference applications and abandoning the general-purpose character.

In this paper, we experimentally present the first photonic AI accelerator that is powered by a microcomb laser source and is capable of executing Matrix-by-Tensor-Multiply (MbTM) operations at a record-high computational power of 262 TOPS. The proposed architecture consists of a 16 × 16 AWGR module, broadband intensity modulators for weight and input encoding at 32 Gbaud, and a Si_3N_4 frequency comb laser source for multi-wavelength generation. This approach extends our recent study on the first AWGR-based PNNs,[36] demonstrating a more than 60% increase in total computational power performance and an ~8.5% higher classification accuracy by exploiting 32 G clock rates and microcomb-laser-generated wavelength channels. The topology relies on the cyclic wavelength routing properties of the AWGR and supports simultaneously time-, wavelength-, and space-division multiplexing (TWSDM) for weight and input vector encoding, forming a powerful framework for matrix and tensor multiplications. To validate the performance of the proposed tensor accelerator in both FC and CNN layouts, two Deep Learning (DL) models were trained for executing different applications while taking into account the optics-informed DL training framework[4] to adapt to the underlying constraints of the photonic hardware, like noise,[13,37] quantization,[38,39] and value non-negativity.[40] The first application was the identification of distributed denial of service (DDoS) attacks in a fully connected NN (FCNN), where the accelerator achieved an experimental Cohen's kappa score of 0.8652 for the 2048 inferred samples, closely matching the software performance. The second task concerned the classification of handwritten digit images (MNIST) through a convolutional NN (CNN), with the hardware inference revealing an accuracy of 92.14%, showing only 1.55% degradation compared to the accuracy achieved via the software.

II. MATRIX-BY-TENSOR MULTIPLICATOR CONCEPT AND PRINCIPLE OF OPERATION

The proposed matrix-by-tensor multiplication engine in its generic N × K × S arrangement is illustrated in the conceptual layout of Fig. 1(a). It comprises an $N \times N$ AWGR module with N and K broadband modulators connected to its #N input and #K output ports (K <= N), respectively. A multi-λ stream, generated by multiplexing N laser beams, is split into N channels of equal powers, with the ith channel (i ∈ [1, N]) entering the ith broadband modulator. The modulator is electrically driven by a weight row vector $\overrightarrow{W_i}(t)$, which contains L elements and is optically imprinted on all #N wavelengths, so that the output of each ith modulator consists of an L-symbol long time series. This implies that the supported weight matrices W can have N rows and L columns, with every weight row vector W_i consisting of L discrete symbols that correspond to the number of discrete time slots allocated for its representation. The AWGR collects the outputs of all i modulators at its input ports and allows all #N different weight row vectors $\overrightarrow{W_1}(t)$, …, $\overrightarrow{W_N}(t)$ to emerge at each AWGR output port by taking advantage of its wavelength cyclic routing properties. Consequently, the whole $N \times L$ weight matrix W appears at every jth output ($j \in [1, K]$), with every weight row vector carried by a different wavelength within the multi-λ stream that emerges at the same AWGR output port.

09 December 2025 17:52:39

© Author(s) 2025

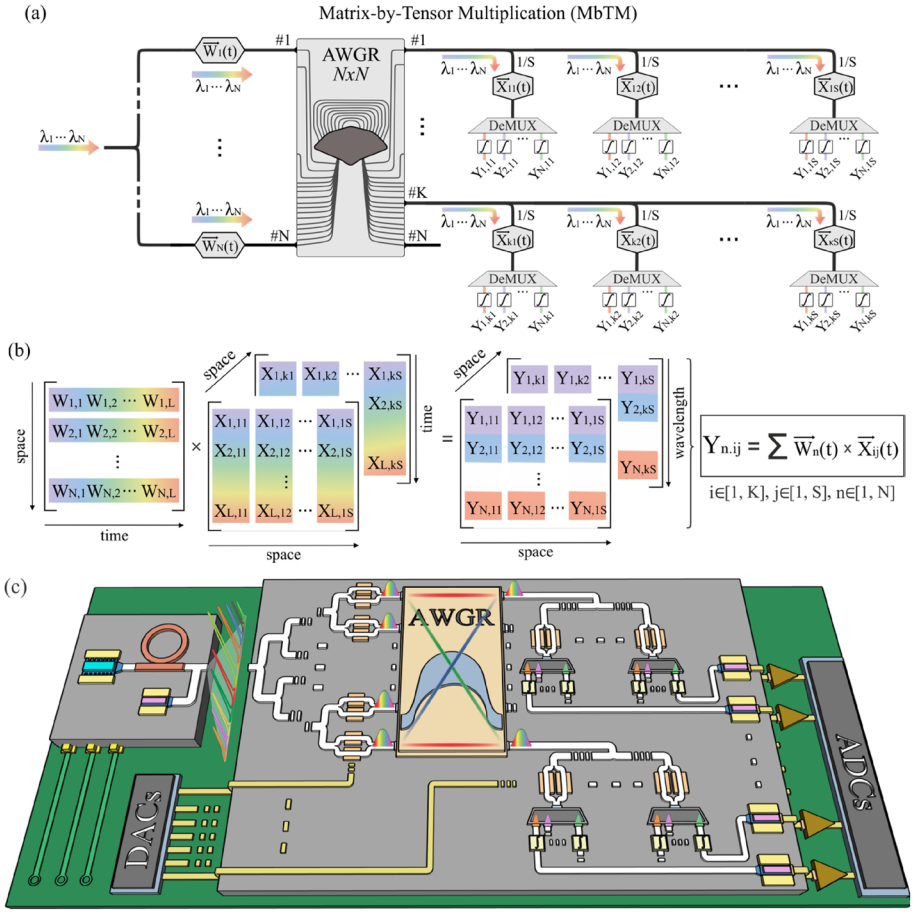


FIG. 1. (a) Conceptual layout of the AWGR-based matrix-by-tensor-multiplication, (b) the corresponding W-matrix and X-tensor with their respective results Y_{ij}, and (c) the envisioned integrated AWGR-based accelerator.

A splitter with a *1:S* ratio at every *jth* AWGR output port is then employed to broadcast the multi-λ weight matrix into *S spatially separated* copies of equal power level. The resulting signals propagate through a X_{jl} broadband modulator that is placed at every splitter output, with $l \in [1, S]$. The $\vec{X_{jl}}(t)$ electrical input column vectors drive the X_{jl}-modulators so that the time instances match the respective *L*-symbol length of the weight vectors. Therefore, the multi-λ signal emerging at the output of the X_{jl}-modulator carries the Hadamard product (HP) $\vec{W_i}(t) \circ \vec{X_{jl}}(t)$ between the $\vec{X_{jl}}(t)$ input time vector and all weight time vectors. By demultiplexing the multi-λ signal arising at the output of the X_{jl}-modulator into its wavelength-components, one obtains every *L*-symbol long HP between a single weight row vector and the input vector. Accumulation is then performed by employing an optical or optoelectronic integrator at every demultiplexer output, which integrates over an *L*-symbol time duration to provide the dot-product, $Y_{i,jl}$, between the vectors $\vec{W_i}(t)$ and $\vec{X_{jl}}(t)$, as proposed in Ref. 41. In this way, a different MbMM, between the weight matrix W and the respective input matrix formed by the *S* different $\vec{X_{jl}}(t)$ column vectors, is calculated for each *j*th AWGR output, as shown in Fig. 1(b). By exploiting the spatial dimension and equipping all *K* AWGR output ports with a similar 1:S split-and-modulation MbMM stage, the proposed architecture can successfully execute a number of *K* different MbMM operations, effectively turning the setup into a MbTM layout where an $N \times L$ weight matrix gets multiplied by an $L \times S \times K$ input signal tensor. In this way, assuming a typical case where $K = N$ for the AWGR input and output ports, the proposed layout supports a total number of $N^2 \cdot S$ computations. By additionally assuming an operational baud-rate of *B* Gbaud at every modulator and a splitting ratio *S* that equals the number of ports *N*, then the number of computations scales as $O(N^3)$, although the circuit complexity scales as $O(N^2)$, resulting in a total computational power of $N^3 \cdot B$ GMAC/sec or, equivalently, $2 \cdot N^3 \cdot B$ GOPS. Figure 1(c) pictorially represents the envisioned AWGR-based architecture, showcasing all key building blocks comprising the proposed topology. In addition, we can define the bandwidth of the parameter load rate as $C_w \times q \times B$, where C_w is the concurrent weight channels (or active weight-imprinting modulators), *q* is the bits/symbol, and *B* remains the baud-rate. For detailed, numerical examples, please see the supplementary material, Sec. S4.

III. EXPERIMENTAL IMPLEMENTATION OF THE AWGR-BASED ACCELERATOR

The experimental implementation of the AWGR-based photonic MbTM engine is illustrated in Fig. 2(a). A silicon nitride

09 December 2025 17:52:39

© Author(s) 2025

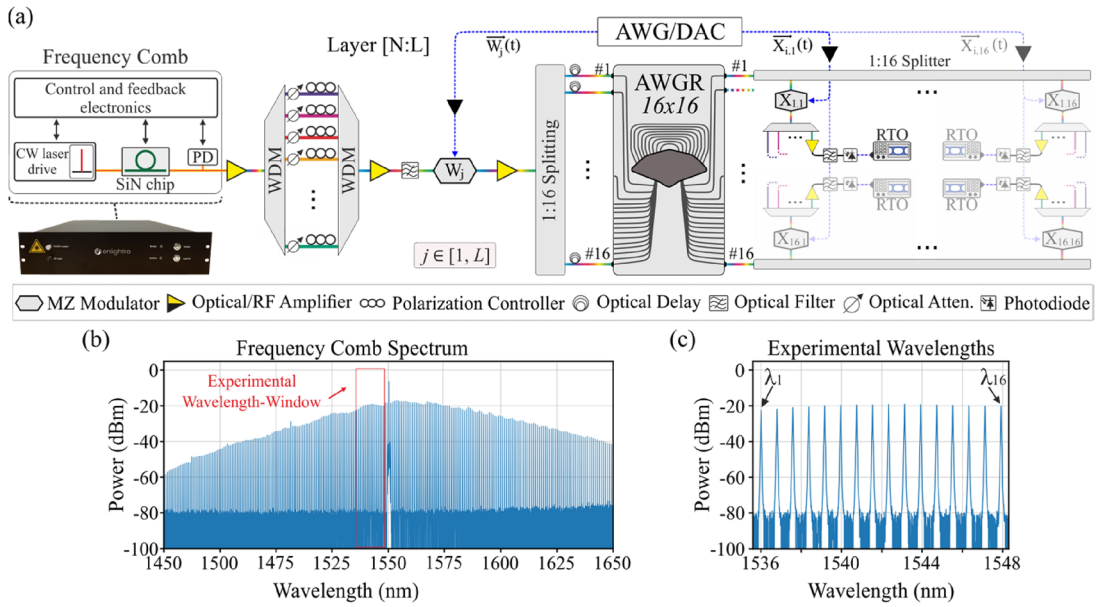


FIG. 2. (a) AWGR-based experimental testbed for MbTM products verification, (b) frequency comb output spectrum, with the red rectangle denoting the experimental wavelength-window, and (c) zoomed-in spectrum of the experimental wavelengths.

(Si_3N_4)-based frequency comb source (Enlightra's SLC) was employed as a continuous wave (CW) generator, with the presented block diagram revealing the basic building blocks: (i) a CW pump laser, (ii) a high-quality Si_3N_4 micro-resonator, and (iii) a controlling unit for the electronics. Light coupled into the ring cavity builds in intensity with each roundtrip pass until the circulating optical power exceeds the material's nonlinear threshold to generate a wide range of frequencies, i.e., a frequency comb, with a free spectral range (FSR) of 100 GHz. The combination of high Kerr nonlinearity in silicon nitride, group velocity dispersion, and a high-Q factor around 1×10^6 of the microresonator enables the generation of a highly coherent optical frequency comb through cascaded four-wave mixing processes. The generated optical spectrum is centered at 1550 nm and features a 3-dB bandwidth of >2 THz with an OSNR reaching 50 dB. An integrated feedback loop with built-in photodiodes (PD on the figure) and electronics ensures long-term stabilization of the system, with >2000 h of non-stop operational stability having been confirmed experimentally. An external graphical user interface allows control of the internal components and parameters of the frequency comb. Enlightra's SLC was emitting a total optical power of 1 mW for all the produced wavelengths, with an erbium doped fiber amplifier (EDFA) connected at the output of the device boosting the optical power. The CW generation stage was followed by a demultiplexing-multiplexing stage allowing for a per-channel adjustment of the selected wavelengths' power levels, ensuring the generation of 16 power-equalized channels with a channel spacing of 0.8 nm within the spectral window of 1535.9–1547.8 nm, with a spacing error of ±0.01 nm. The polarization state of every channel was also adjusted on a per-channel basis to allow for efficient amplitude modulation of all 16 wavelengths in the subsequent modulation stage. A second EDFA was placed after the multiplexer for loss compensation, with a bandpass filter (BPF) of 12 nm optical bandwidth removing the amplified spontaneous emission noise. The filtered multi-λ stream with a total average power of 28.1 mW, or ~1.8 mW per wavelength-channel, was then injected into an indium phosphide (InP) Mach–Zehnder modulator (MZM) with 35 GHz electro-optical (EO) bandwidth (Fraunhofer HHI InP MZM 64 Gbaud), which was driven by an arbitrary waveform generator (AWG – Keysight M8194A) to generate the $\overrightarrow{W_j}(t)$ row vector of the W-matrix. The AWG generated the electrical signal with a power of 350 mV, which was further amplified via an RF amplifier (SHF M804B), reaching ~3.5 V as necessitated from the InP MZM. In this way, an L-value weight row vector $\overrightarrow{W_j}(t)$ was imprinted as a time series onto all 16 wavelength-channels. After optically amplifying the W-modulator output to compensate for losses in the subsequent components, the signal was split into 16 identical spatial components via a 1:16 splitting stage, with the 16 WDM signals getting decorrelated via the employment of different fiber-length delays at every of the 16 paths prior to entering the respective AWGR input. The employed AWGR (Semicon SAWG-G-100G-32-32-C) is commercially available and consists of 32 input and 32 output ports with a 3 dB bandwidth of 0.4 nm. For this experiment though, we only exploited the 16 input and 16 output ports acting in this way as a 16 × 16 AWGR. As such, the $\overrightarrow{W_j}(t)$ row vector entering the AWGR via a certain input port will emerge at every one of the 16 AWGR output ports but at a different wavelength at each port due to the cyclic wavelength routing properties. This means that every AWGR output port will carry 16 different wavelengths, with every wavelength stemming from a different AWGR input and carrying a time-delayed copy of the $\overrightarrow{W_j}(t)$ row vector. In this way, only one wavelength can be considered at a certain AWGR

09 December 2025 17:52:39

© Author(s) 2025

output to carry the targeted $\overrightarrow{W_j}(t)$ row vector, with the rest of the 15 wavelength-channels exiting through the same AWGR output, aggregating the channel under evaluation. The multi-λ modulated stream derived at the kth AWGR output was further split by a 1:16 splitter, creating 16 identical copies at respective paths. Each ith splitter output port was fed to a $LiNbO_3$ X_{ki}-input MZM (iXBlue MX-LN-40) with 40 GHz EO bandwidth, designated to modulate the corresponding $\overrightarrow{X_{ki}}(t)$ input vector. The $LiNbO_3$ MZM was driven again by the AWG module to produce the optical X_{ki} input vector time series, and as before, a RF amplifier (SHF L806A) was employed to amplify the 200-mV electrical signal generated via the AWG. By sequentially connecting the X_{ki}-input MZM to the 16 splitter outputs of the kth AWGR output port, we successfully acquire the MbMM product. After repeating the above-mentioned procedure for the remaining AWGR ports by successively connecting the 1:16 split-and-modulate AWGR output stage to all AWGR outputs, the required MbTM products are obtained. To evaluate the different Hadamard products carried by every wavelength, the multi-λ signal was demultiplexed into its 16-wavelength constituents at the output of every X_{ki}-MZM. Every channel was then amplified in an EDFA followed by a BPF with 0.55 nm bandwidth prior to entering a 70 GHz photodiode (PD, Finisar XPDV3120) and being recorded by a 256 GSa real-time oscilloscope (RTO, Keysight UXR0704AP). A software-based filter was implemented via the RTO and applied to the captured signal, with a manually adjusted 3 dB bandwidth of 20 or 32 GHz in order to mitigate the excess noise bandwidth of the PD. The integration and non-linear activation function (AF) were performed off-line via a respective software routine.

IV. EXPERIMENTAL VALIDATION IN DL APPLICATIONS

The experimental validation of the proposed MbTM architecture was carried out in two different AI applications that exploit two different DL models and respective NN configurations. The scope of the two applications was (i) cybersecurity, through the identification of DDoS attacks in Data Centers (DCs) via the analysis of data packet traffic and the classification between malicious and benign packets, and (ii) classification of handwritten digits, using the MNIST dataset. The training procedure for the DDoS identification and MNIST classification tasks started with the software-based implementation of the task-specific NN models, i.e., an FC for the DDoS and a CNN for the MNIST, using the PyTorch framework. In parallel, following the pipeline described in Ref. 38, we extracted the noise of the hardware, which was modeled based on the std, for each targeted operational speed. As also described in the supplementary material, Sec. S1, following this procedure enabled the extraction of the noise-equivalent bit resolution (NEB) for our photonic hardware, which later defined the training constraints.

The first step was to apply a linear dimensionality reduction to the DDoS dataset using Principal Component Analysis (PCA) to reduce the input vector size of the telemetry data to 6. The extraction process of the characteristics for the DDoS dataset is further detailed in Ref. 42. After each layer, the digital sigmoid AF provided the non-linearity of the system. The NN was trained with the Adam optimizer[43] for 100 epochs, applying hard constraints to weights, forcing them to non-negative values within the range of [0, 1]. These weight constraints were incorporated during training by introducing a penalty term to the loss function, formulated as

$$L(\hat{y}, y; w) = L_c(\hat{y}, y) + \alpha \sum_{k=0}^{K} \sum_{i=0}^{N_k} \sum_{j=0}^{M_k} \frac{\sqrt{\max\left\{-w_{ij}^{(k)}, 0, w_{ij}^{(k)} - 1\right\}^2}}{N_k M_k}, \tag{1}$$

where L_c denotes the cross-entropy loss and α is the weighting factor for the penalty term, which by default is set to $\alpha = 10$. The N_k and M_k define the fan-in and fan-out of the kth layer, respectively. In addition, the hard constraints for the clipping of the weights in the range of [0, 1] were applied after the tenth epoch, following the optimization step.

The implementation of the CNN for MNIST classification employed three convolutional software-based layers, two convolutional hardware-based layers, and one fully connected hardware-based layer. The software-based convolutional layers introduced a backbone for feature extraction and consisted of three layers. The first two (Conv 1, 2) employed a 3 × 3 kernel with three channels each and applied the digital ReLU AF as the non-linearity. The third (Conv 3) employed a 3 × 3 kernel with four channels, with a stride and dilation equal to 2, with the non-linearity implemented by the digital sigmoid AF. The backbone's output, a four-channel 11 × 11 feature map, was fed into the cascaded layers, which were validated through the optical hardware. The first optical convolutional layer (Conv 4) applied a 4 × 4 kernel with eight channels and a stride of 4. The extracted feature map continued to the second optical convolutional layer (Conv 5) that applied a 2 × 2 kernel with 32 channels. The output of the last photonic convolutional layer (Conv 5) was flattened to a column vector consisting of 32 values ready to be processed by the optical classification/output layer (FC). The non-linearity selected for the hardware-based layers was experimentally extracted by a programmable opto-electro-optical system with the operational principles detailed in Ref. 44. The CNN was trained in an end-to-end manner for 75 epochs, applying different optimization approaches to the software-backbone layers and to the hardware layer. Specifically, the backbone optimization employed the Adam optimizer with a learning rate equal to 0.001 and included a learning rate scheduler to reduce its value by half every 25 epochs. The optical layers were restricted to 3-bit precision and employed a normalized quantization-aware training approach.[39] Optimization of the optical layers utilized the multiplicative Adam optimizer[40] to ensure that the non-negativity applied during weight initialization remained during the training process. In this way, the hard constraints on the optical layers' weights were mitigated using only the upper-bound penalty term of the applied loss function of Eq. (1), with $\alpha = 1$.

Figure 3(a) illustrates an instance of telemetry data of the data traffic within a DC, which were used in the DDoS identification task. The telemetry data were categorized into six classes, corresponding to the six input vectors of the first FC NN layer, as shown in Fig. 3(b), followed by an output FC layer and two outputs. Each of the neurons of layer 1 (L1) and layer 2 (L2) is described by the layout of Fig. 3(c). Prior to evaluation, a preliminary pre-emphasis procedure (discussed in the supplementary material) was employed to compensate for (i) the noise originating from the limited frequency response of the deployed modulators and (ii) the non-linearities within the electro-optic system. In order to allow the processor architecture to utilize all its weighting and input signal modulators for useful computations without sacrificing any wavelength or modulator resources for negative number representation,[45] the DDoS classification NN was trained to allow only for non-negative

09 December 2025 17:52:39

© Author(s) 2025

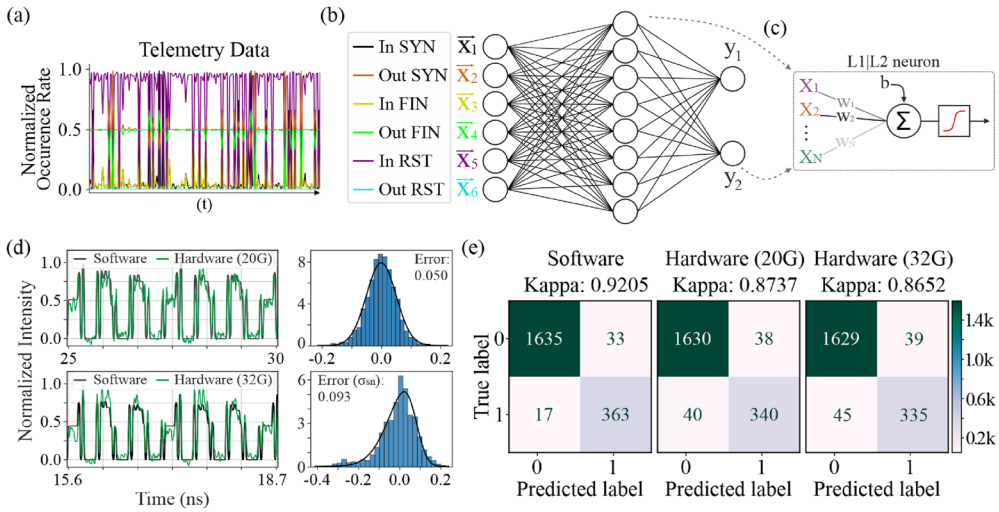


FIG. 3. (a) Telemetry data of the DC traffic, (b) fully connected NN topology, (c) L1 and L2 neuron breakdown, (d) time traces for the software experimental traces at 20 and 32 Gbaud, with their respective errors, and (e) confusion matrices obtained from the inference performed on software, hardware at 20 and 32 Gbaud, along with the respective achieved kappa scores.

values through the use of optics-informed DL models.[4] Figure 3(d) illustrates the software-obtained data together with the experimental time traces of the Hadamard products $\overrightarrow{W_i}(t) \circ \overrightarrow{X_{jl}}(t)$ obtained at the first splitter output connected to the first AWGR output at wavelength λ_1 at two different symbol rates, i.e., 20 and 32 Gbaud, respectively. As can be observed, the experimental traces closely follow their respective software-obtained counterparts. The noise distribution at both 20 and 32 Gbaud is also

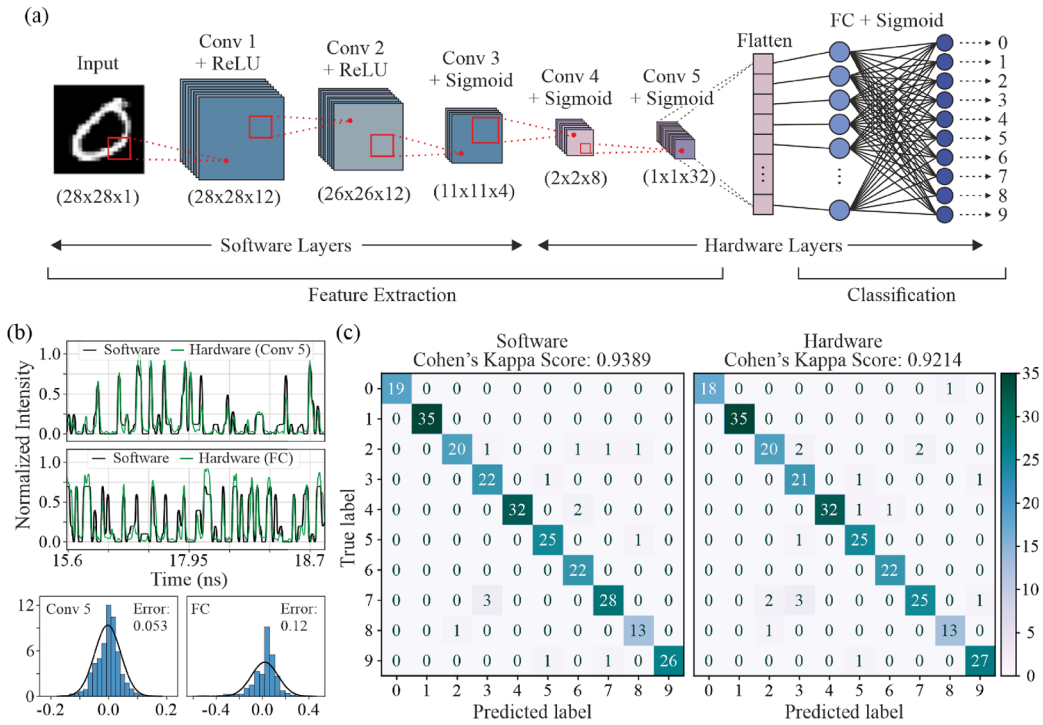


FIG. 4. (a) The CNN network used for MNIST classification, with the software and hardware layers being denoted; (b) software and hardware traces at 32 Gbaud for Conv 5 and FC layers, together with the respective errors; and (c) confusion matrices for software and hardware inferences, with their kappa scores.

09 December 2025 17:52:39

© Author(s) 2025

shown in Fig. 3(d), revealing a standard deviation (STD) of only 0.05 when fitted with a Gaussian distribution for 20 Gbaud, while for the error of the 32 Gbaud operation, the Azzalini skew-normal fit[46] was employed to also consider the negative tail of the histogram bins, exhibiting a σ_{SN} = 0.093. Similar results were obtained for all wavelengths across all 16 outputs of the split-and-modulate stage at every AWGR output port. The DDoS attack recognition used a set of 2048 input samples consisting of 81.45% (1668) benign and 18.55% (380) malicious packets, which resulted in a highly imbalanced dataset. The Cohen's kappa-score[27,47] metric accounts for the imbalance of classes and, therefore, provided a more accurate representation in the final validation. The confusion matrices resulting from the software-based inference and the hardware when operating at 20 Gbaud and then at 32 Gbaud are presented in Fig. 3(e), along with their respective kappa scores. The scores obtained with this hardware, i.e., 0.8718 and 0.8677 for the 20- and 32-Gbaud cases, respectively, indicate excellent performance with minor degradation compared to the software case.

To evaluate the performance of the photonic AI accelerator also in CNNs, a second experimental validation process was carried out for the classification of handwritten digits through a hybrid software/hardware NN. More specifically, the NN topology comprised five convolutional layers and a fully connected layer for the final classification. Among these layers, the first three convolutional layers were executed in software and were responsible for the initial dimensionality reduction, while the remaining two convolutional layers and the last fully connected layer were implemented over the photonic hardware. The NN topology and the respective size of each layer are illustrated in Fig. 4(a). Again, an optics-informed DL training scheme was employed to allow for the use of strictly non-negative values and take into account the different value, quantization, and noise constraints of the photonic hardware. Figure 4(b) depicts an indicative time trace at the output of the CNN layer five (Conv 5) and the FC layer for the software and hardware multiplications, after the software correlation, for a time-window of 4.7 ns, or 150 symbols, at 32 Gbaud. The experimental traces closely follow those acquired from the software, with similar results being obtained for all different wavelengths across all possible PNN outputs. The noise distribution of all output waveforms is shown in the same figure and reveals a noise STD of only 0.053 and 0.12 for the Conv 5 and FC layer, respectively. This NN task classified a total of 256 images using software and hardware inference with relatively balanced subsets for each class. The confusion matrices of the software- and hardware-based classifications are depicted in Fig. 4(c). The software-based inference for the selected 256 samples reached 94.53% accuracy, and the hardware-based inference, at 32 Gbaud operation, achieved 92.57% accuracy. This difference corresponds to a misclassification of only five samples for the hardware inference. The kappa-score values were also calculated to account for the slightly imbalanced classes in the dataset, revealing software- and hardware-based values of 0.9389 and 0.9214, respectively.

V. DISCUSSION

The proposed AWGR-based multidimensional tensor-multiplication demonstrator consisted of a 16 × 16 AWGR module, a frequency comb laser providing the 16 carrier wavelengths, and high-speed MZMs. By driving the weight and input signal modulators at 20 Gbaud, the total computational power reaches the value of 163.84 TOPS, which increases to the record-high value of 262 TOPS when increasing the symbol rate at 32 Gbaud, exhibiting a ~60% increase in computational power when compared to our previous study[36] or a ~2200% increase when compared to the next in the exhibited TOPS, the PNN architecture.[23] Figure 5(a) juxtaposes the recent PNN demonstrators in terms of the achieved compute performance in TOPS, where a projection of a future implementation of a 32 × 32 AWGR accelerator is included. Following the same rationale, Fig. 5(b) depicts the comparison of our envisioned SiPho-integrated demonstrators with the established electronic state-of-the-art accelerators.[48–61] It should be noted that the energy efficiency of our architecture follows the calculations presented in the supplementary material, Sec. S3.

The demonstrated version of the hyperdimensional AWGR-based hardware prototype employed fiber-based components but can potentially be transferred to a chip-scale integrated version, taking into account the current capabilities of silicon photonic integration technology. A viable roadmap in view of transferring our prototype architecture to an integrated form is shaped along exploiting the developments of the optical interconnect infrastructure

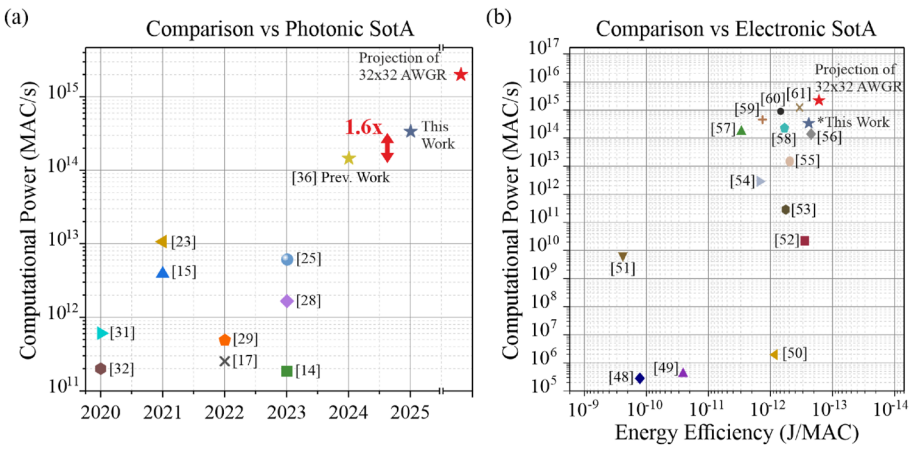


FIG. 5. Comparison of state-of-the-art. (a) Photonic and (b) electronic topologies. *Projected EE of this study, based on SiPho components (see the supplementary material, S3), for the 16 × 16-AWGR accelerator and for 32G operation.

09 December 2025 17:52:39

© Author(s) 2025

toward leveraging the low-cost and high-volume credentials offered by SiPho technology. In the current implementation, the transmitter (T×) and receiver (Rx) arrays are established with bulky fiber-based components; hence, the first step would focus on migrating to readily available deployments of SiPho T×,[62] R×,[63] and T×/R× array prototypes.[64] The next step would target replacing the currently employed silica-based (de)MUX and AWGR devices with SiPho-based integrated prototypes, capable of reducing the architecture's chip real-estate requirements. High-port-count demonstrators have already shown their credentials in Si-based[65,66] and SiN-based[67] platforms, exhibiting ultra-compact 32× (de)MUX[65] and high-performing 16 × 16 AWGR[66] prototypes, while enabling low-loss and low-fabrication variation characteristics toward employing low-phase noise with reduced crosstalk AWGR modules.[67] The final stage of integration envisions a multi-chiplet architecture,[68] interconnected through an electro-optic printed circuit board (PCB)[69] capable of hosting both electronic and photonic subsystems. Moving to a chiplet-based integration provides a key advantage, as it eliminates the area and thermal restrictions inherent to monolithic, single-reticle processors. In this scheme, distinct functionalities are separated, each undertaken by a dedicated chiplet: (i) a laser chiplet generates the multi-wavelength optical carriers, (ii) a modulator chiplet implements the input and weight encoding modules forming the computational core, (iii) the AWGR can be integrated on another chiplet along with the (de)MUX and photodiode arrays, and (iv) a final electronic chiplet hosts the DAC/ADC and driver circuitry. This modular organization serves two main purposes. First, it enables scaling to larger systems by combining multiple smaller T×/R× chiplets. Second, it permits cross-platform compatibility, allowing components from different silicon photonics platforms to be assembled into a single system. Overall, this roadmap builds on established optical interconnect technologies and current trends in photonic integration, projecting that the hardware cost trajectory will continue to follow the downward curve characteristic of photonic interconnects.[70] While this concept could represent a potential path toward realizing large-scale AWGR-based tensor accelerators, the present discussion serves as a preliminary outlook, and a detailed exploration of such multi-chiplet implementations remains beyond the scope of this study. Nevertheless, we have already initiated efforts in this direction, which could enable full-scale realizations of the proposed architecture.[71]

Although setting an integration roadmap could benefit the prospect of transferring our prototype to a SiPho integrated platform, the scalability of the 1:N and 1:S splitting ratios is limited. To estimate the maximum N, S ratios, we first need to set the requirements in laser power and amplification for the proposed accelerator, which will consequently determine (i) the number of active components and (ii) their respective power consumption. These numbers can be calculated based on the IL_{Total} (see the supplementary material, S3). The starting point of this analysis is the sensitivity of the receiver circuitry (R_{sens}), which for the envisioned SiPho AWGR-based accelerator topology is set to −16 dBm at 20 Gbaud,[72] which is in line with state-of-the-art transimpedance amplifier (TIA) performance. The IL values included in the analysis were chosen based on the literature as IL_{WDM} = 1.6 dB,[65] IL_{mod} = 3.7 dB,[73] and IL_{AWGR} = 3 dB.[67] For each $N \times N$-sized AWGR topology (N = 4,8,16,32), the IL of each splitting stage can be calculated based on $IL_S = 3 \times log_2(N)$. Hence, the higher the scale of the topology, the more

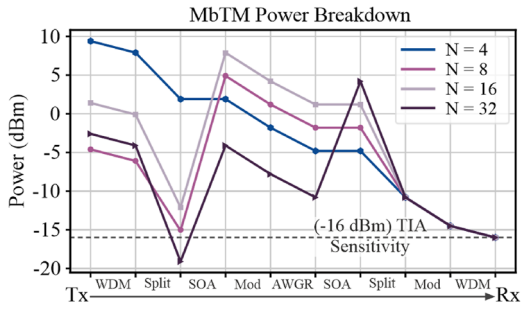


FIG. 6. Feasibility analysis for the $N \times N$-AWGR based MbTM photonic accelerator, with N = 4,8,16,32, considering the insertion loss (IL) of each stage and the respective amplification requirements.

IL introduced from the N, S splitting ratios, which in turn necessitates the introduction of amplification stages. However, to calculate the N, S and the amplifications, we should account for the optical power limitation within the waveguide, prior to which nonlinear effects dominate, which is assumed to be ∼16 dBm for Si-based waveguides.[74] Applying these limitations, a breakdown analysis was conducted for the proposed system, considering the less-possible active components within the system. The emission power of each comb-line was assumed based on the targeted value of the next generation and optimized for the Enlightra SLC frequency comb. The Si_3N_4 microresonator technology allows for selection of a narrower band, e.g., 16 or 32 lines, which can further boost the per-line optical power to 10 mW, while careful engineering enables a smaller per-line fluctuation, at ±2 dB. The laser power reflects a single line of the optimized comb, wherein each case was aptly selected to match the limitations of intra-waveguide power while also considering the sensitivity of the receiver. Figure 6 illustrates the integration-feasibility analysis of the proposed MbTM demonstrator, indicating the need for optical amplification at N = 8,16,32, with a 20 dB gain amplifier incorporated for the cases of N = 8,16. For the case of N = 32, the increased IL, owed to the higher splitting ratio, necessitates two 15 dB gain amplification stages. This analysis constitutes a first indication for the feasibility of the proposed hyperdimensional tensor-accelerator while also reflecting on the maximum N, S splitting ratios, defined directly from the inherent linear-power within the Si-waveguides.

Figure 7 depicts the scalability perspectives of the architecture and plots the two key metrics of computing engines, i.e., computational power (CP) and energy efficiency (EE) vs the AWGR size, considering state-of-the-art integrated photonic components (supplementary material, S3). The analysis assumes the use of an integrated $N \times N$ AWGR, while the PNN operates at three different data rates: 20 and 32 Gbaud, as presented in this study, and a future target of 50 Gbaud that has already been shown to successfully support AI applications at chip-scale.[27] For the feasibility of integrating the whole system, the aggregated in-band crosstalk of the AWGR and (de)multiplexing devices should also be considered. Crosstalk can be modeled as additive noise and bounded relative to the effective bit precision of the system. A detailed calculation of the crosstalk requirements together with closed-form scaling limits is provided in the supplementary material, Sec. S5. Quantification of the respective EE requires a detailed breakdown of the power

09 December 2025 17:52:39

© Author(s) 2025

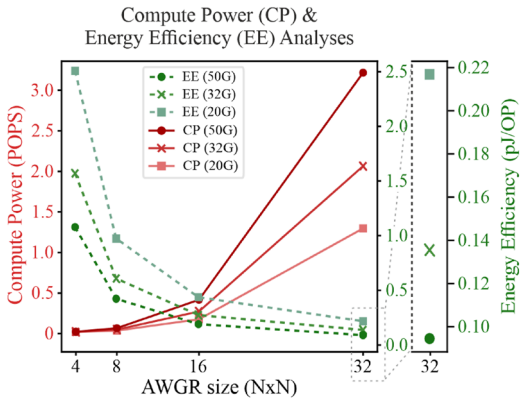


FIG. 7. Scalability analysis of compute power and energy efficiency vs AWGR size. 1 POPS = 10^{15} ops/s.

consumption (PC) of all its active components, incorporating also the electronic digital-to-analog conversion (DAC), integrator, and analog-to-digital conversion (ADC) circuits. The assumptions about the speed and energy consumption of a chip-scale PNN version are provided in detail in the supplementary material. The MbTM throughput scales with $O(N^3)$, and the achieved operations can be calculated as $2 \cdot N^3 \cdot B$ GOPS, for K=N, with a more detailed analysis of the throughput being presented in the supplementary material, Sec. S4. As such, the total computational power increases with the cube of the AWGR dimensions, while energy efficiency improves since the total power consumption scales with the total number of active components employed, which in turn scales with $O(N^2)$. The computational power and energy efficiency also improve with increasing data-rate, indicating that a silicon photonic version of the 16 × 16 AWGR-based PNN layout at 32 Gbaud allows for a total compute power of 262 TOPS with an energy efficiency of 273 fJ/OP. Assuming a 32 × 32 AWGR[75] silicon-based PNN operating at 50 Gbaud symbol rates could then, in principle, enable a total computational power up to $2 \cdot 32^3 \cdot 50 \cdot 10^9 = 3.276$ POPS, within a power consumption envelope of 309.5 W, suggesting an energy efficiency of ~94 fJ/OP. Although the estimated absolute power consumption of ~310 W is higher than that of present-day electronic accelerators, the true value of the proposed architecture lies in its throughput-per-watt scaling. While the computational power scales as $O(N^3)$, the power consumption grows only as $O(N^2)$, leading to an overall ~2× improvement in energy-per-operation. This becomes evident when doubling the I/O port count of the AWGR, where the computational power scales by a factor of 8× while the power consumption increases by only ~4×. Such scaling behavior constitutes the key advantage of our AWGR-based tensor engine, while the further maturation of the SiPho technology is expected to further improve modulation efficiency, AWGR IL, and co-packaged integration, which can further reduce the absolute consumption, thereby narrowing the gap with electronics while preserving the scaling advantage.

VI. CONCLUSIONS

We have experimentally validated an MbTM architecture that comprises a comb laser source, a 16 × 16 AWGR, and broadband MZMs driven up to 32 Gbaud data-rates, providing a record-high computational power of 262 TOPs. Two different DL models for respective applications were trained and utilized for the experimental validation of the proposed MbTM layout in the AI domain: a FC NN for DDoS identification and a deep NN formed by multiple convolutional layers and a last FC layer for MNIST classification. Successful classification between benign and malicious traffic in the DDoS attack identification task was experimentally observed, with an experimentally obtained Cohen's kappa-score of 0.8677 over 2048 inference samples with only a 0.05 degradation compared with software. The classification of digits within the MNIST dataset achieved an experimental accuracy of 92.14% over the hardware inference of 256 samples, which closely matched the software performance of 93.89%. Finally, we discussed the integration perspectives of the proposed architecture and highlighted this as a promising roadmap toward additional improvements in computational power and energy efficiency. The projected AWGR-based architecture, with upgraded connectivity to 32 × 32 and a compute rate of 50 Gbaud, is expected to allow for 3.276 POPS of computational power consuming sub-100 fJ/OP, which constitutes a ~298× increase in computational power compared to state-of-the-art photonic accelerators.

SUPPLEMENTARY MATERIAL

The supplementary material provides additional details on the pre-emphasis process for the photonic link linearization, validation of multi-wavelength operation, power consumption breakdown, computational throughput analysis, crosstalk tolerance, and an illustrative discussion on the compatibility of the AWGR-based accelerator with GEMM algorithms for contextual benchmarking. It also contains the referenced figures and tables (supplementary material, Figs. 1–2 and Tables 1–2) that support the results and analysis presented in the main paper.

ACKNOWLEDGMENTS

Enlightra would like to acknowledge Charlotte Bost for her study during PIC linear characterization and Lou Kanger for nonlinear characterization. Enlightra would also like to acknowledge Alexey Feofanov for his assistance in the development of the Enlightra SLC. The study was, in part, funded by the Chips Joint Undertaking project HAETAE (Grant No. 101194393) and Horizon 2020 projects Gatepost (Grant No. 101120938) and ALLEGRO (Grant No. 101092766).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

C.P., T.M., M.M.P., A.T., and N.P. conceived the experiment. C.P., A.P., and T.M. deployed the experimental setup, performed the experiment, and processed the experimental results. M.K., O.A., N.P., and A.T. performed the training of the neural network models. C.O. manufactured the device and packaged the photonic integrated circuit. T.S. simulated the design of the photonic integrated circuit and developed the software for the device. A.Y.T., J.D.J., and M.K.

09 December 2025 17:52:39

© Author(s) 2025

supervised the study. C.P., M.K., A.T., and N.P. wrote the paper. All authors discussed the results.

Christos Pappas: Conceptualization (supporting); Data curation (equal); Formal analysis (lead); Methodology (equal); Validation (equal); Writing – original draft (lead); Writing – review & editing (lead). **Antonios Prapas**: Data curation (equal); Formal analysis (supporting); Software (lead); Validation (supporting). **Theodoros Moschos**: Conceptualization (supporting); Data curation (equal); Formal analysis (supporting); Methodology (supporting). **Manos Kirtas**: Writing – original draft (supporting). **Odysseas Asimopoulos**: Software (supporting); Visualization (supporting). **Apostolos Tsakyridis**: Conceptualization (equal); Methodology (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Miltiadis Moralis-Pegios**: Conceptualization (equal); Methodology (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Chris Vagionas**: Conceptualization (supporting); Supervision (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Nikolaos Passalis**: Supervision (supporting); Writing – original draft (supporting). **Cagri Ozdilek**: Data curation (supporting); Resources (supporting); Writing – original draft (supporting). **Timofey Shpakovsky**: Resources (supporting). **Alain Yuji Takabayashi**: Writing – original draft (supporting). **John D. Jost**: Resources (equal). **Maxim Karpov**: Resources (equal). **Anastasios Tefas**: Conceptualization (supporting); Software (supporting); Supervision (supporting); Writing – original draft (supporting). **Nikos Pleros**: Conceptualization (lead); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

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

REFERENCES

[1] J. L. Hennessy and D. A. Patterson, "A new golden age for computer architecture," Commun. ACM **62**(2), 48–60 (2019).
[2] D. Amodei and D. Hernandez (2018). "AI and compute," OpenAI, https://openai.com/index/ai-and-compute/ (accessed 13 Sep. 2025)
[3] J. Sevilla *et al.*, "Compute trends across three eras of machine learning," arXiv:2202.05924 (2022).
[4] A. Tsakyridis *et al.*, "Photonic neural networks and optics-informed deep learning fundamentals," APL Photonics **9**(1), 011102 (2024).
[5] B. J. Shastri, A. N. Tait, T. Ferreira de Lima *et al.*, "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics **15**, 102–114 (2021).
[6] Z. Zhong *et al.*, "Lightning: A reconfigurable photonic-electronic SmartNIC for fast and energy-efficient inference," in *Proceedings of the ACM SIGCOMM 2023 Conference* (ACM Conferences, 2023), pp. 452–472.
[7] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**, 441–446 (2017).
[8] A. Totovic *et al.*, "WDM equipped universal linear optics for programmable neuromorphic photonic processors," Neuromorph. Comput. Eng. **2**, 024010 (2022).
[9] I. Roumpos *et al.*, "High-performance end-to-end deep learning IM/DD link using optics-informed neural networks," Opt. Express **31**(12), 20068 (2023).
[10] C. Pappas *et al.*, "A TeraFLOP photonic matrix multiplier using time-space-wavelength multiplexed AWGR-based architectures," in *Optical Fiber Communications (OFC)* (IEEE, San Diego, CA, 2024), pp. 1–3.
[11] C. Pappas *et al.*, "A 160 TOPS multi-dimensional AWGR-based accelerator for deep learning," in *2024 Optical Fiber Communications Conference and Exhibition (OFC)* (IEEE, San Diego, CA, 2024), pp. 1–3.
[12] R. Hamerly *et al.*, "Large-scale optical neural networks based on photoelectric multiplication," Phys. Rev. X **9**, 021032 (2019).
[13] G. Mourgias-Alexandris *et al.*, "Noise-resilient and high-speed deep learning with coherent silicon photonics," Nat. Commun. **13**, 5572 (2022).
[14] G. Giamougiannis *et al.*, "Analog nanophotonic computing going practical: Silicon photonic deep learning engines for tiled optical matrix multiplication with dynamic precision," Nanophotonics **12**(5), 963–973 (2023).
[15] J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**, 52–58 (2021).
[16] H. Zhang *et al.*, "An optical neural chip for implementing complex-valued neural network," Nat. Commun. **12**, 457 (2021).
[17] F. Ashtiani *et al.*, "An on-chip photonic deep neural network for image classification," Nature **606**, 501–506 (2022).
[18] G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, "An all-optical neuron with sigmoid activation function," Opt. Express **27**, 9620 (2019).
[19] S. Kovaios *et al.*, "Programmable Tanh- and ELU-based photonic neurons in optics-informed neural networks," J. Lightwave Technol. **42**(10), 3652–3660 (2024).
[20] Z. Chen *et al.*, "Deep learning with coherent VCSEL neural networks," Nat. Photonics **17**, 723–730 (2023).
[21] M. Moralis-Pegios *et al.*, "Neuromorphic silicon photonics and hardware-aware deep learning for high-speed inference," J. Lightwave Technol. **40**(10), 3243–3254 (2022).
[22] G. Giamougiannis *et al.*, "Universal linear optics revisited: New perspectives for neuromorphic computing with silicon photonics," IEEE J. Sel. Top. Quantum Electron. **29**(2), 6200116 (2023).
[23] X. Xu *et al.*, "11 TOPS photonic convolutional accelerator for optical neural networks," Nature **589**, 44–51 (2021).
[24] J. Cheng *et al.*, "Direct optical convolution computing based on arrayed waveguide grating router," Laser Photonics Rev. **18**, 2301221 (2024).
[25] E. Luan *et al.*, "Towards a high-density photonic tensor core enabled by intensity-modulated microrings and photonic wire bonding," Sci. Rep. **13**, 1260 (2023).
[26] A. N. Tait, T. F. de Lima, E. Zhou *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**, 7430 (2017).
[27] G. Giamougiannis *et al.*, "Neuromorphic silicon photonics with 50 GHz tiled matrix multiplication for deep-learning applications," Proc. SPIE **5**(1), 016004 (2023).
[28] B. Dong *et al.*, "Higher-dimensional processing using a photonic tensor core with continuous-time data," Nat. Photonics **17**, 1080–1088 (2023).
[29] A. Sludds *et al.*, "Delocalized photonic deep learning on the internet's edge," Science **378**, 270–276 (2022).
[30] A. Tsakyridis *et al.*, "Universal linear optics for ultra-fast neuromorphic silicon photonics towards Fj/MAC and TMAC/sec/mm^2 engines," IEEE J. Sel. Top. Quantum Electron. **28**(6), 8300815 (2022).
[31] 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).
[32] G. Giamougiannis *et al.*, "A coherent photonic crossbar for scalable universal linear optics," J. Lightwave Technol. **41**(8), 2425–2442 (2023).
[33] T. Zhou *et al.*, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," Nat. Photonics **15**, 367–373 (2021).
[34] H. H. Zhu *et al.*, "Space-efficient optical computing with an integrated chip diffractive neural network," Nat. Commun. **13**, 1044 (2022).
[35] Z. Xu *et al.*, "Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence," Science **384**, 202–209 (2024).
[36] C. Pappas *et al.*, "Reaching the peta-computing: 163.8 TOPS through multidimensional AWGR-based accelerators," J. Lightwave Technol. **43**(4), 1773–1785 (2025).

09 December 2025 17:52:39

© Author(s) 2025

[37]M. Moralis-Pegios *et al.*, "Neuromorphic silicon photonics and hardware-aware deep learning for high-speed inference," J. Lightwave Technol. **40**(10), 3243–3254 (2022).
[38]M. Kirtas *et al.*, "Robust architecture-agnostic and noise resilient training of photonic deep learning models," IEEE Trans. Emerging Top. Comput. Intell. **7**(1), 140–149 (2023).
[39]M. Kirtas *et al.*, "Quantization-aware training for low precision photonic neural networks," Neural Networks **155**, 561–573 (2022).
[40]M. Kirtas, N. Passalis, and A. Tefas, "Multiplicative update rules for accelerating deep learning training and increasing robustness," Neurocomputing **576**, 127352 (2024).
[41]N. Youngblood, "Coherent photonic crossbar arrays for large-scale matrix-matrix multiplication," IEEE J. Sel. Top. Quantum Electron. **29**, 6100211 (2023).
[42]M. Kirtas *et al.*, "Early detection of ddos attacks using photonic neural networks," in *IEEE 14th IVMSP Workshop* (IEEE, 2022), pp. 1–5.
[43]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980v9 (2017).
[44]C. Pappas *et al.*, "Programmable Tanh-ELU-Sigmoid-and sin-based nonlinear activation functions for neuromorphic photonics," IEEE J. Sel. Top. Quantum Electron. **29**, 6101210 (2023) Photonic Signal Processing.
[45]F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, "CrossLight: A cross-layer optimized silicon photonic neural network accelerator," in *2021 58th ACM/IEEE Design Automation Conference (DAC)* (IEEE, San Francisco, CA, 2021), pp. 1069–1074.
[46]A. Azzalini, *The Skew-Normal and Related Families* (Cambridge University Press, 2013).
[47]S. M. Vieira, U. Kaymak, and J. M. C. Sousa, "Cohen's kappa coefficient as a performance measure for feature selection," in *International Conference on Fuzzy Systems* (IEEE, Barcelona, Spain, 2010), pp. 1–8.
[48]B. V. Benjamin *et al.*, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," Proc. IEEE **102**(5), 699–716 (2014).
[49]M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," IEEE Micro **38**(1), 82–99 (January/2018).
[50]N. Jouppi *et al.*, "TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings," in *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA'23)* (Association for Computing Machinery, New York, NY, 2023), pp. 1–14 Article 82.
[51]S. B. Schemmel, P. Dauer, and J. Weis, "Accelerated analog neuromorphic computing," in *Analog Circuits for Machine Learning, Current/Voltage/Temperature Sensors, and High-Speed Communication*, edited by B. Murmann and B. Hoefflinger (Springer, Cham, Switzerland, 2022), pp. 197–225.
[52]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), 1537–1557 (2015).
[53]Z. Jia, B. Tillman, M. Maggioni, and D. P. Scarpazza, "Dissecting the graphcore IPU architecture via microbenchmarking," arXiv:1912.03413 (2019).
[54]NVIDIA Corporation, "NVIDIA Tesla V100 GPU architecture," White Paper, 2017, https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.
[55]Mythic, Inc., "Analog compute-in-memory: architecture and efficiency for edge AI," White Paper, 2019, https://mythic.ai/wp-content/uploads/2022/02/MythicWhitepaper-2019oct31.pdf.
[56]NVIDIA Corporation, NVIDIA Jetson AGX Orin series: AI at the edge, Technical Brief, 2022, https://developer.nvidia.com/embedded/jetson-agx-orin.
[57]AMD, "AMD instinct MI200 series accelerators: High-performance computing and AI," White Paper, 2021, https://www.amd.com/system/files/documents/instinct-mi200-series-whitepaper.pdf.
[58]Y. Wang, C. Li, and C. Zeng, "Exploring the performance bound of cambricon accelerator in end-to-end inference scenario," in *2nd BenchCouncil International Symposium on Benchmarking, Measuring, and Optimizing (Bench 2019), Denver, CO, USA, Nov. 14–16, 2019, Revised Selected Papers* (Springer, Cham, Switzerland, 2020), pp. 67–74.
[59]NVIDIA Corporation, "NVIDIA RTX Blackwell GPU Architecture," White Paper, 2025, https://images.nvidia.com/aem-dam/Solutions/geforce/blackwell/nvidia-rtx-blackwell-gpu-architecture.pdf.
[60]Intel Corporation, "Habana Gaudi3 AI processor architecture," White Paper, 2024, https://www.intel.com/content/dam/www/central-libraries/us/en/documents/habana-gaudi3-ai-processor-architecture-white-paper.pdf.
[61]D. Abts *et al.*, "Think fast: A tensor streaming processor (TSP) for accelerating deep learning workloads," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)* (IEEE, Valencia, Spain, 2020), pp. 145–158.
[62]M. Rakowski *et al.*, "Hybrid 14 nm FinFET—Silicon photonics technology for low-power Tb/s/mm^2 optical I/O," in *2018 IEEE Symposium on VLSI Technology*(IEEE, Honolulu, HI, 2018), pp. 221–222.
[63]P. De Heyn *et al.*, "Ultra-dense 16 × 56Gb/s NRZ GeSi EAM-PD arrays coupled to multicore fiber for short-reach 896 Gb/s optical links," in *Optical Fiber Communication Conference (OFC)* (IEEE, 2017), paper Th1B.7.
[64]M. Moralis-Pegios *et al.*, "4-channel 200 Gb/s WDM O-band silicon photonic transceiver sub-assembly," Opt. Express **28**, 5706–5714 (2020).
[65]Q. Deng *et al.*, "32 × 100 GHz WDM filter based on ultra-compact silicon rings with a high thermal tuning efficiency of 5.85 mW/π," in *Optical Fiber Communication (OFC) Conference* (IEEE, San Diego, CA, 2024), pp. 1–3.
[66]K. Fotiadis *et al.*, "Silicon photonic 16 × 16 cyclic AWGR for DWDM O-Band interconnects," IEEE Photonics Technol. Lett. **32**(19), 1233–1236 (1 Oct.1, 2020).
[67]K. Shang, S. Pathak, C. Qin, and S. J. B. Yoo, "Low-loss compact silicon nitride arrayed waveguide gratings for photonic integrated circuits," IEEE Photonics J. **9**(5), 6601805 (2017).
[68]M. Moralis-Pegios *et al.*, "Silicon circuits for chip-to-chip communications in multi-socket server board interconnects," IET Optoelectron. **15**, 102–110 (2021).
[69]T. Lamprecht *et al.*, "Electronic-photonic board as an integration platform for Tb/s multi-chip optical communication," IET Optoelectron. **15**, 92–101 (2021).
[70]C. Minkenberg *et al.*, "Co-packaged datacenter optics: Opportunities and challenges," IET Optoelectron. **15**, 77–91 (2021).
[71]A. Prapas *et al.*, "Time-space-wavelength multiplexed photonic tensor core using WDM SiGe EAM array chiplets," ' Opt. Express **33**, 36960–36972 (2025).
[72]S. Saeedi, S. Menezo, G. Pares, and A. Emami, "A 25 Gb/s 3D-Integrated CMOS/silicon-photonic receiver for low-power high-sensitivity optical communication," J. Lightwave Technol. **34**, 2924–2933 (2016).
[73]E. Chong *et al.*, "112G+7-Bit DAC-based transmitter in 7-nm FinFET with PAM4/6/8 modulation," IEEE Solid-State Circuits Lett. **5**, 21–24 (2022).
[74]S. Singer *et al.*, "Nonlinear loss and damage threshold in silicon photonic waveguides: Modelling and experimental verification," in *2022 Conference on Lasers and Electro-Optics (CLEO)* (Optica Publishing Group, 2022), p. SF3O.4.
[75]X. Shen, W. Zhao, H. Li, and D. Dai, "High-performance silicon arrayed-waveguide grating (de)multiplexer with 0.4-nm channel spacing," Adv. Photonics Nexus **3**, 036012 (2024).

09 December 2025 17:52:39

© Author(s) 2025