

*Abstract***—The explosion of artificial intelligence and machinelearning algorithms, connected to the exponential growth of the exchanged data, is driving a search for novel application-specific hardware accelerators. Among the many, the photonics field appears to be in the perfect spotlight for this global data explosion, thanks to its almost infinite bandwidth capacity associated with limited energy consumption. In this review, we will overview the major advantages that photonics has over electronics for hardware accelerators, followed by a comparison between the major architectures implemented on Photonics Integrated Circuits (PIC) for both the linear and nonlinear parts of Neural Networks. By the end, we will highlight the main driving forces for the next generation of photonic accelerators, as well as the main limits that must be overcome.**

*Index Terms***—Matrix-vector multiplication, photonics, PICs, silicon photonics, tensor core.**

Fig. 1. Breaking down of a Fully-Connected Neural Network. (a) Example of a NN having one hidden layer. (b) Every single neuron receives the input signals from all the previous layer neurons, scaled by a factor *w*, performing their summation and passing through an activation function. (c) This single neuron can be generalized by including the whole layer, employing a matrix representation. The matrix representation \mathcal{L}

consumption and latency for the reasons, research has reasons, research has reasons, research has reasons, res started to look for novel technologies that can provide a better hardware acceleration for $\frac{1}{\sqrt{2}}$, $\frac{1}{\sqrt{2}}$, raised as an alternative approach for hardware implementation of NN, thanks to its speed-of-light latency and low energy consumption $[19, 20, 21, 22, 23.$ Moreover, Silicon Photonics has started to become a reliable and diffuse technology, allowing the implementation of Photonic Network \mathbf{P} $\left(\begin{array}{c} -1\text{V}^+ & -1\text{V}^- & -1\text{V}^+ & -1\text{V}^+ & -1\text{V}^+ \end{array}\right)$ needs of final users

ASIC can be found in many companies, such as Nvidia, Intel, and Tesla

Fig. 3. Scaling comparison between the original Reck scheme $[66]$, and the original $[66]$, and the original $[66]$ one proposed by Feng et al. 72 , the connections proposed to reduce the connections 33 .

On-chip Fiber Way 19:3.

 $\mathcal{F}(\mathcal{S}, \mathcal{F}_{\mathcal{A}}) \leq \mathcal{F}_{\mathcal{A}} \leq \mathcal{F}_{\mathcal{A}}$

the potential to leverage on the different types of material that \mathcal{A}_k can be used, limited by the compatibility with the compatibility with the SiPh CMOS of the SiPh CMOS of the Si
The SiPh CMOS of the SiPh process. The main drawback can be identified in the scaling limitations: since the input power must meet a certain criteria to active the nonlinear function $\left(\begin{array}{cc} -1 & -1 \\ -1 & 0 \end{array}\right)$ and the propagation $\left(\begin{array}{cc} 1 & -1 \\ -1 & 0 \end{array}\right)$ is still affected by dB/cm losses, a large NN with several consecutive layers, or a high loss PIC would not be suitable for this approach, unless other adjustments (like on-chip amplification on-chip amplification and $\binom{1}{1}$ stage) would be adopted.

C. Light Splitting and Detection

Another approach has been used by Moayedi et al. $\frac{103}{103}$, and B_1 , shown respectively in Fig. 3(c) [\(e\).](#page-7-0) In this case, the linear part of the NN is based on the MZI mesh, and the nonlinear function is activated by just part of the light power of the output waveguide (splitting), which is detected and \mathbf{r} the signal used to modulate the amplitude of the amplitude of the remaining particle \mathcal{F}_{max} of the optical signal. This allows feeding the whole network with $\frac{1}{\sqrt{2}}$ the same optical input signals, reducing the need to have more lasers or input couplers. However, the network will add layers of modulation on top of each other, making the same scheme more sensitive to noise and not directly suitable for WDM expansions. Following the same approach, $\frac{5}{2}$ et al. $\frac{114}{2}$ propose a similar s scheme. In this case the NL part is implemented using a MZI, where one of the arm is controlled by a optical memory-based by a optical memory-based by a optical memory-base
The arm is controlled by a optical memory-based by a optical memory-based by a optical memory-based by a serie feedback circuit, using a PCM material as nonvolatile element. The light-splitting-and-detection has some clear advantages as the modulation is directly on the same optical signal, with a clear advantage in terms of speed and latency. However, the taper of speed and latency.
However, the taper of the taper requires an electrical circuit capable of reading low currents from the reading low currents from the currents f
The currents from the currents from the currents from the currents from the current from the current from the the photodetectors and translate into proper signals, limiting $\left\{ \begin{array}{c} \mathbf{t} & \mathbf{t} \\ \mathbf{t} & \mathbf{t} \end{array} \right\}$ energy consumption as well, determine by the Transimpedance amplifiers (TIA). Moreover, the continuous splitting layer after a function \mathbb{R}^n layer increases the insertion loss of the insertion loss of the insertion loss of the overall photonic circuit,
In the overall photonic circuit, and insert photonic circuit, and insert photonic circuit, and instruct constr putting more pressure on the performances of the latter activation functions in terms of minimum optical power detections of minimum optical power detection.
المستحدث الأمريكي المستحدث ال In all cases, the activation function function function $\mathcal{E}(\mathcal{A})$ is encoded at the hardware $\mathcal{E}(\mathcal{A})$ $\mathcal{N} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad \mathcal{N} = \sum_{i=1}^N \mathcal{$ outputs, and so fixing the scalability and limiting parallelization. Schemes that can be used to subdivide the matrix into smaller ones to fit Ω on smaller hardware cannot be used by u with hardware activation functions, as the non-linear functions, as the nonis a priori, and so it does not allow for a temporal so it does not allow for a temporal so it does not allow for a temporal solution \mathcal{L} multiplexing summation. By so, the research may investigate $\frac{1}{2}$ schemes that allow an actual flexible implementation of the nonlinear function, by exploiting more programmable photonic photonic programmable photonic photonic photonic p
The photonic photoni

steps will focus on inter-chip communications, as well as \overrightarrow{Domain} Crossings:

- *Domain Crossings:* Photonic-based tensor core processors are analog in nature and hence may require digital to analog and vise versa domain crossings. Above $5G$ \mathcal{F} baud rates and 8-bit resolution DACs and ADCs become quite experiments of $\overline{55}$. If the PTC application allows a probability $\overline{55}$. processing data in the optical domain (from an optical in the intra data-center, for interaction $\frac{1}{2}$, for example, for example, for example, $\frac{1}{2}$ photonic PIC-based DAC would be beneficial [\[125\].](#page-12-0) This could include also energy harvesting, such as recapturing optical nonlinearities (126) , nanoscale RF antennas or solar antennas or sol \sim cells \sim 127].
- \blacksquare *Architectures:* \blacksquare , see the set proposed and demonstrated. While a clear winner is still to be found, all of them can push towards several improvements to further expanding the $\frac{23}{5}$. One side, $\frac{23}{5}$. parallelization exploiting of freedom can further degrees of freedom can further degrees of further degrees of furthe push the performances. On the performances of the other side, techniques. On the other side, techniques, t
The other side, techniques, techniques, techniques, techniques, techniques, techniques, techniques, techniques such as pruning or others can be implemented on-chip as well, making room for improvements in the overall system.

- 125 J. M. \ldots \ldots "Electronic bottleneck suppression in next-generation networks with in t_{rel} converted photonic digital-to-analog converters, $\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{2}$ $2, 2021, \ldots$ 2000033.
- $[126 \quad B_{\text{max}}, 0]$ is radius, and $[126 \quad B_{\text{max}}]$ $\mathbb{Z}^{\{1\}}$ **Photonics**, *IEEE J. Sel. Topics Quantum Electron.*, ¹, ¹12, $\frac{1}{2}$. 3, $\frac{1}{2}$. 412–421, MJ . 2006.
- $\begin{bmatrix}127\end{bmatrix}$ M. H. T. Sorger, "Enhanced photon absorption in spiral in spira $\sum_{i=1}^n$ $\sum_{i=1}^n$ $\sum_{i=1}^n$ $\sum_{i=1}^n$ $\sum_{i=1}^n$ $\sum_{i=1}^n$ $\sum_{i=1}^n$ *Nanotechnol.*, $\frac{1}{2}$, 26, .34, 2015, $\frac{1}{2}$, .344005.
- 128 Ω , \ldots , \ldots , \ldots , Ω , \ldots , \ldots \mathcal{G} , \mathcal{F} , \mathcal{G} is the design space of a configurable space of a configuration of a co hybrid nanophotonics, *Microprocess. Microsyst.*, ¹, 50, \ldots 113–126, 2017.
- $\begin{bmatrix}129\end{bmatrix}$ C. Shen et al., $\begin{bmatrix}109\end{bmatrix}$ configurable application-specific photonic in- $\frac{|1|^2}{\sinh 2208.03588} \frac{|1|^2}{(1+1)^{1/2}} \frac{|1|}{(1+1)^{1/2}} \frac{|1|^{1/2}}{|1+1|^{1/2}} \frac{|1|^{1/2}}{|1+1|^{1/2}} \frac{|1|^{1/2}}{|1+1|^{1/2}}$ *arXiv:2208.03588*.
- $\frac{130}{\pi} \int_{0}^{\frac{\pi}{2}} \left(45 + \frac{1}{2}\right)^{12} \int_{0}^{\frac{\pi}{2}} \frac{45}{\pi} \int_{0}^{\frac{\pi}{2}} \left(45 + \frac{1}{2}\right)^{12} \left(45 + \frac{1}{2}\right) \int_{0}^{\frac{\pi}{2}} \left(45 + \frac{1}{2}\right)^{12} \left(45 + \frac{1}{2}\right) \int_{0}^{\frac{\pi}{2}} \left(45 + \frac{1}{2}\right) \left(45 + \frac{1}{2}\right) \left(45 + \frac{1}{2}\right) \int_{0$ nology (45CL) for $\mathbf{f} = \mathbf{f} \cdot \mathbf{g}$ and high speed optical speed optical speed optical speed optical speed *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2020, pp. 1–3.
- [131] W.-T. J. Chan, Y. Du, A. B. Kahng, S. Nath, and K. Samadi, "BEOL stack-aware routability prediction from placement using data mining t_{max} in *Proc. IEEE 34th Int. Conf. Comput. Des.*, 2016, pp. 41–48.
- 132 T. Standaert et al., $\frac{1}{2}$ nm technology integration for the 7 nm technology integration for the 7 nm technology in ¹, *Proc. IEEE Int. Interconnect Technol. Conf./Adv. Metallization Conf.*, 2016, $\frac{2}{5}$, $\frac{2}{5}$
- $\begin{bmatrix}133\end{bmatrix}$ $\begin{bmatrix}0\end{bmatrix}$ $\begin{bmatrix}0\end{$ \mathbb{R} , \mathbb{R} in and self-aligned-block options for the $\frac{1}{3}$ nm node, $\frac{1}{3}$ in *Proc. IEEE Int. Electron Devices Meeting*, 2019, pp. 19–3.
- 134 M. R. $\frac{1}{2}$ M. R. Billah et al., $\frac{1}{2}$ integration of silicon photonics circuits and photonical photonics and photonics and photonics circuits and photonics and photonics and photonics and photonics and phot I_1 lastronic wire bonding, photonic wire bonding, I_2 , I_3 , I_4 , I_5 , I_7 , I_8 , I_8 , I_8 , I_9 , $2018.$

Nicola Peserico $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 & 1 \\$ $s=\frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{2}}\right)$ and $\frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{2}}\right)$ and $\frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{2}}\right)$ ment of \mathbf{G} and $\mathbf{$ Washington, DC, USA. His research interests include silicon photonics, AI/ML accelerations, optoelectronics devices and components, and components and components, and bio-sensing with with photonic integrated circuits.

Bhavin J. Shastri(Senior Member, IEEE) received the Ph.D. degree in electrical received the Ph.D. de engineering (photonics) from $G_{\text{tr}} = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{2} \sum_{i=1}^{n} \frac{1}{2}$ 2012. He is currently an Assistant Professor of $P_{\rm eff}$ 1.1247 , $5(^{-1})$, $\frac{1}{2}$ $\frac{20126}{\sqrt{2}}$ (x $\frac{1}{2}$ S and a Faculty Affiliate with the Vector $\frac{1}{2}$ $\Psi_{\alpha\beta}$ institute for Artificial Intelligence, $\mathbb{X}_{\alpha\beta}$ See and Banting 2016–2018 and Banting NSERC Postdoctoral Fellow during \mathcal{N}_{S} Post $2012^{11}2016$ with Princeton University, Princeton, $P_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ or coauthored more than 70° journal articles and 100 conference processes and 100 conference processes 70° seven book chapters, and given more than 65 invited talks and lectures, five and lectures, five keynotes and four tutorials. He is a coauthor of the book (CRC Press, 2017) Neuromorphic P a term he helped coin. H research interests include silicon photonics, photonic integrated circuits, neuromorphic computing, and machine learning. Dr. Shastri was the recipient of the 2022 $\frac{1}{2}$ and the 2020 $\frac{1}{2}$ and $\frac{2020}{2}$ IUPAP Young Scientist Prize in Optics for Optics f his pioneering contributions to neuromorphic photonics from ICO. He is a Senior Member of Optica (formerly Oscar of the recipient of the recipient of the 2014 \mathbf{G} Postmanting Felix from the Government of Canada, the 2012 \mathcal{L} because \mathcal{L} and \mathcal{L} \mathcal{L} and \mathcal{L} at \mathcal{L} student at McGill, and \mathcal{L} $\frac{a^2}{a^2} + \frac{1}{a^2} + \frac{a^2}{a^2} + \frac{a^2}{a^2} + \frac{a^2}{a^2} + \frac{a^2}{a^2} + \cdots + \frac{1}{a^2} + \frac{a^2}{a^2} + \cdots$ awards.

a Fellow of the German National Academic Foundation. He is a \mathbf{G} \mathbf{G} \mathbf{G} \mathbf{H} \mathbf{G} \mathbf{G} \mathbf{H} \mathbf{G} \mathbf{G} \mathbf{H} \mathbf{G} \mathbf{G} \mathbf{H} \mathbf{H} \mathbf{G} \mathbf{H} \mathbf{H} \mathbf{H} Optelligence.