Perspective

Frank Brückerhoff-Plückelmann, Ivonne Bente, Daniel Wendland, Johannes Feldmann, C. David Wright, Harish Bhaskaran and Wolfram Pernice*

A large scale photonic matrix processor enabled by charge accumulation

https://doi.org/10.1515/nanoph-2022-0441
Received July 31, 2022; accepted October 17, 2022; published online October 28, 2022

Abstract: Integrated neuromorphic photonic circuits aim to power complex artificial neural networks (ANNs) in an energy and time efficient way by exploiting the large bandwidth and the low loss of photonic structures. However, scaling photonic circuits to match the requirements of modern ANNs still remains challenging. In this perspective, we give an overview over the usual sizes of matrices processed in ANNs and compare them with the capability of existing photonic matrix processors. To address shortcomings of existing architectures, we propose a time multiplexed matrix processing scheme which virtually increases the size of a physical photonic crossbar array without requiring any additional electrical post-processing. We investigate the underlying process of time multiplexed incoherent optical accumulation and achieve accumulation accuracy of 98.9% with 1 ns pulses. Assuming state of the art active components and a reasonable crossbar array size, this processor architecture would enable matrix vector multiplications with $16{,}000 \times 64$ matrices all optically on an estimated area of 51.2 mm^2, while performing more than 110 trillion multiply and accumulate operations per second.

***Corresponding author: Wolfram Pernice**, Department of Physics, Heidelberg University, Kirchhoff-Institute for Physics, Im Neuenheimer Feld 227, 69120 Heidelberg, Germany, E-mail: wolfram.pernice@kip.uni-heidelberg.de. https://orcid.org/0000-0003-4569-4213
Frank Brückerhoff-Plückelmann, Department of Physics, University of Münster, CeNTech, Heisenberg Str. 11, 48155 Muenster, Germany
Ivonne Bente, Department of Physics, University of Münster, CeNTech, Heisenberg 11, 48149 Muenster, Germany. https://orcid.org/0000-0003-3198-4244
Daniel Wendland, Department of Physics, University of Münster, CeNTech, Heisenberg 11, 48149 Muenster, Germany
Johannes Feldmann, Salience Labs Ltd, 46 Woodstock Rd, Oxford OX2 6HT, UK
C. David Wright, University of Exeter, Faculty of Environment, Science and Economy, North Park Road, Exeter, UK
Harish Bhaskaran, University of Oxford, Department of Materials, Parks Road, Oxford OX1 3PH, UK

Keywords: matrix vector multiplication; photonic computing; time-multiplexing.

1 Introduction

Matrix vector multiplications (MVMs) are the computational backbone of artificial neural networks (ANNs) as they mathematically describe the connections between neurons in the layers the network is composed of. Therefore, energy efficient, compact and high-speed matrix processors are crucial to power the complex ANNs deployed for autonomous driving [1] and language processing [2] among other important applications. Integrated photonic circuits are an attractive approach to implementing MVM processors due to their large optical bandwidth [3–6], inherent low latency [7] and minimal heating losses in comparison to electronic approaches. Prototypes deploying Mach–Zehnder interferometer (MZI) meshes [8], photonic crossbar arrays (PCAs) [9] and ring resonator weight banks [10] have demonstrated that photonic computing is viable and highlight the capabilities for MVM operations. However, one of the largest functional systems built to date is a 64×64 matrix processor deploying MZI meshes which is, even in view of being a substantial achievement, still of moderate size considering that modern ANNs consist of billions of free parameters. This particular system occupies a chip area of 150 mm^2 [11]. Substantially increasing the size of such photonic circuits is challenging due to fabrication imperfections impacting the splitting ratios in the MZI meshes/crossbar arrays and the overall optical loss of the system. Therefore, a range of architectures and approaches are being developed which could enable processing larger ANNs in the photonic domain.

In this perspective, we give an overview on ANNs deployed in computer vision, natural language processing and combinatorial optimization and compare their requirements with the capability of different integrated photonic matrix processors. As a particular addition

Open Access. © 2022 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

building on photonic processors, we propose a time multiplexed architecture that employs a high-speed reconfigurable photonic crossbar array to allow for MVM processing of larger matrices. Time multiplexing virtually increases the size of the PCA without suffering from the drawbacks of having to fabricate large photonic circuits. Moreover, this approach does not require any additional electrical processing. We experimentally characterize the process of time multiplexed incoherent optical accumulation and propose a design of the photonic circuit underlying the required matrix processor and estimate its performance.

2 Integrated photonics for artificial neural networks

On a basic level, artificial neural networks consist of linear matrix vector multiplications and non-linear activation functions. Additionally, their physical implementation requires memory and a process/data flow corresponding to the chosen ANN architecture. Integrated photonics promise to speed up ANNs in an energy efficient manner in two different ways. First, the photonic circuit computes the full ANN directly in one or few processing steps [12, 13]. Second, the photonic circuit only computes a part of the ANN, for example the MVMs [8–11, 14, 15]. Only performing the MVMs drastically reduces the complexity of the photonic circuit and hence allows accelerating more complex ANNs, but this advantage comes at the cost of having to perform more electro-optic conversions.

Figure 1A shows the typical input vector sizes of the MVM processors depending on the application. Convolutional neural networks perform excellently in computer vision. The main building blocks are convolutional layers which calculate the convolution of the layer input with a trained kernel. The employed kernels are typically rather small, ranging from 3×3 to 11×11 for many networks [16–19], which in turn leads to small input vector sizes for one channel. In contrast, transformer architectures excel in the area of natural language processing. Their main feature is the attention module, which computes correlations between the different symbols in the input sequence. The input vector size of the performed MVMs depends on the model dimension which is often on the order of 10^3 [20–23]. Finally, recurrent neural networks like Hopfield nets and Boltzmann machines can find the ground state of the Ising model and hence are suitable to obtaining good solutions of NP-hard problems like the travelling salesman problem [24–27]. The input vector size of the underlying MVMs directly translates to the number of spins in the ising model, where state of the art ising machines can simulate spins in the order of 10^4 to 10^6 [28–31].

In contrast, Figure 1B shows the (projected) input vector size of various integrated photonic matrix processors. In photonic computing, the input vector is encoded in optical pulses and the respective multiplications are carried out via interaction of the pulses with phase-shifters, attenuators, or amplifiers [8, 32, 33]. There are two different approaches to accumulate the weighted optical pulses, performing either a coherent, phase sensitive superposition [8] or performing an incoherent superposition [9]. While coherent superposition also enables subtraction via destructive interference, practical systems require phase error compensation. In contrast, when using incoherent signals, two pulses of different wavelength are temporally overlapped to perform an incoherent superposition. Since the result of the MVM is calculated at the photodetectors, all interference effects are averaged out if the frequency detuning between both pulses is larger than the detector bandwidth [34]. The main drawback of incoherent superposition is that subtraction cannot be performed optically. The advantage is that the overall system is much more tolerant to fabrication imperfections and measurement conditions because of the inherent phase insensitivity. Furthermore, incoherent superposition can be easily used in combination with wavelength division multiplexing (WDM) [35]. However, both approaches are using a physical photonic circuit to represent the matrix vector multiplication leading to huge circuits for large scale MVMs. A common workaround is to deploy tiled matrix multiplication to reduce the matrix size; however this creates additional electronical overhead. In this perspective we propose an architecture that virtually increases the size of the photonic matrix processor based on time multiplexed incoherent optical accumulation. This allows computing MVMs with large input vector size all optically (in contrast to electronic addition of tiles, which require much more analogue to digital conversions and additional electronic circuitry) on a chip size compatible with commercial foundry processes.

3 Time multiplexed incoherent optical accumulation

In order to virtually enlarge the processing capacity of photonic MVM processors, we combine the concept of incoherent superposition with charge accumulation inside

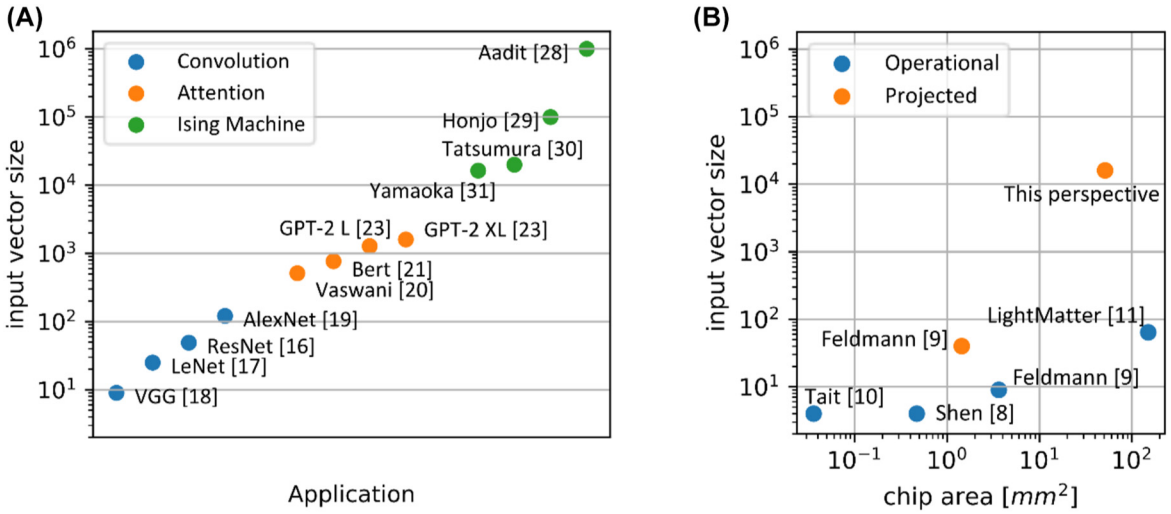

Figure 1: Applications and prototypes of (optical) neural networks. (A) Typical input vector size of the MVMs for different applications. Convolutional layers convolve the input with a kernel, typically in the range of 3×3 to 11×11 for computer vision. The attention head is the main block of the transformer architecture excelling in natural language processing. The input vector sizes depend on the model dimension, usually in the order of 10^3. Hopfield and Boltzmann networks can mimic the Ising model and thus find good solutions to complex computational problems. The input vector size directly translates to the number of spins, which is in the range of 10^4 to 10^6 for modern Ising machines. (B) Input vector size of operational and projected integrated photonic matrix processors.

a photodetector (Figure 2A). We perform incoherent superposition by temporally overlapping two pulses of different wavelengths and accumulate several pulses of the same wavelength by making use of charge accumulation inside the photodetector [36]. If the delay between the pulses is short in comparison to the inverse detector bandwidth, the detector cannot distinguish between the individual pulses. Instead, the output signal of the detector has higher amplitude. To show the feasibility of this approach experimentally, we use 1 ns pulses at 1550 and 1560 nm together with a 10 MHz detector (New Focus Model 2053) to characterize the process of time multiplexed incoherent optical accumulation. Figure 2B shows the detector output signal of two delayed pulses normalized to the manually summed output signal of the individual pulses for different delays between the pulses. We perform each individual measurement 10 times, where the measurement setup itself shows a noise level in the order of 1–2%. Up to a delay of 10 ns, the error caused by charge accumulation inside the detector is within the uncertainty of the measurement setup itself. Next, we perform time accumulation simultaneously with incoherent superposition. We accumulate four pulses of two different wavelengths with a time delay of 1.86 ns in each wavelength channel. Figure 2C shows the measured sum of the four pulses versus the calculated sum of the pulses. The measured signals are distributed around the ideal line with a standard deviation of 1.1% which is again inside the uncertainty of the measurement itself. The main advantage of this approach is the scaling properties. Individual approaches, incoherent superposition and charge accumulation, scale linearly with the number of wavelength channels/pulses. However, the combination of both scales with the number of wavelength channels × number of pulses. Moreover, the comparably slow detector and readout process enables low noise operation which in combination with high saturation powers makes scaling possible (in our experiments, the deployed detector has a noise equivalent power of 0.34 pW/Hz and the saturation power is 10 mW). For broadband operation, the wavelength dependent responsivity of the detector can be compensated by setting a wavelength dependent calibration factor.

4 Large-scale photonic matrix vector multiplication

Since the accumulation is independent from the multiplication for incoherent crossbar arrays, we can directly transfer the concept of time multiplexed optical accumulation to the complete system. In this way, we virtually increase the size of a photonic crossbar array beyond its physical dimensions without inducing any additional electrical processing. Figure 3A depicts how the matrix multiplication $y = M \cdot x$ is performed. In the context of

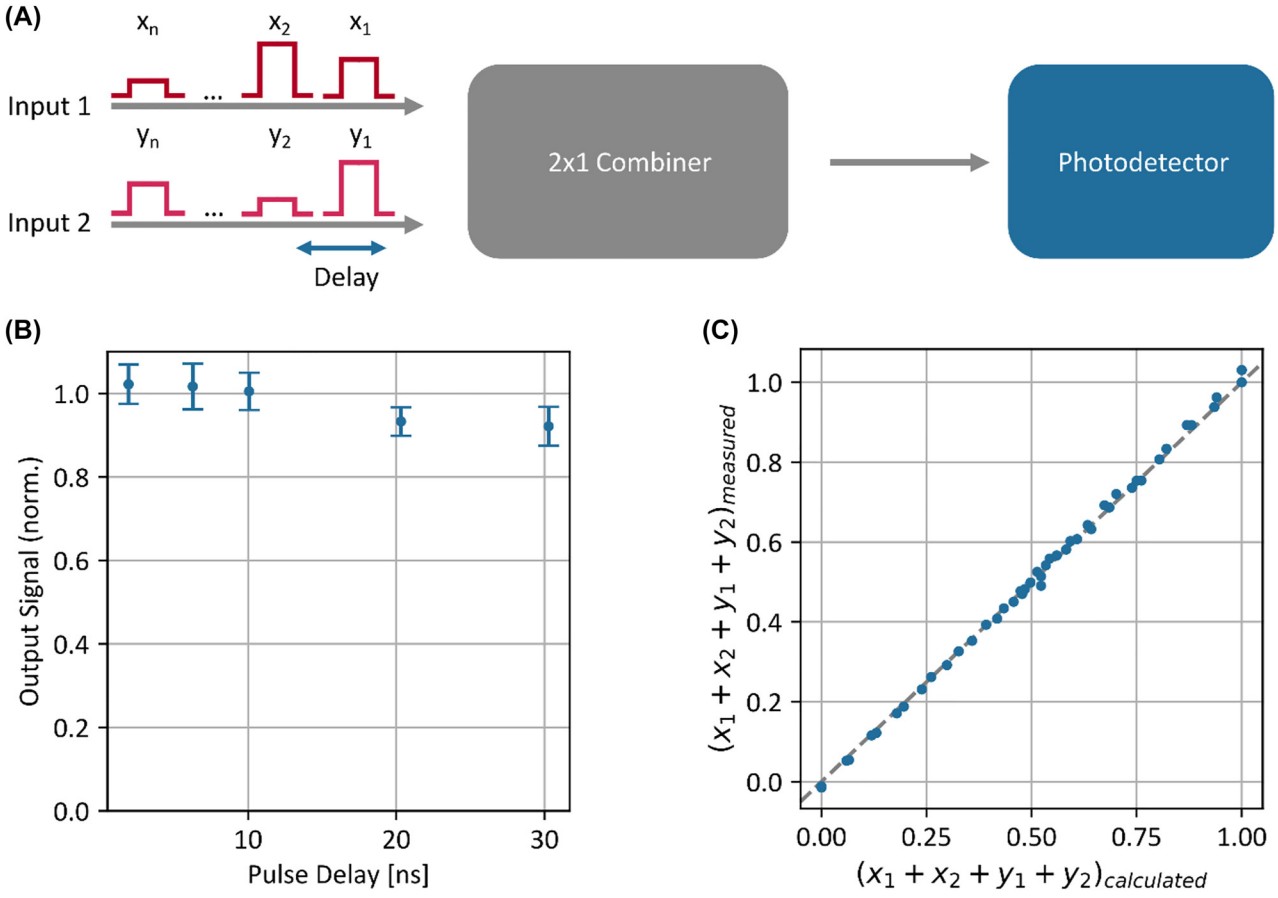

Figure 2: Time multiplexed incoherent accumulation. (A) Concept of time multiplexed incoherent accumulation. We combine two accumulation schemes, temporally overlapping pulses of different wavelength and charge accumulation inside a detector by delayed pulses. Both processes are phase insensitive. (B) Accumulation time of a 10 MHz Photodetector. We determine the accumulation time of the photodetector by comparing the detector signal of the delayed accumulated pulse with the signal of the individual pulses. Up to a delay of about 10 ns the error induced by this accumulation scheme is within the uncertainty of the measurement setup. (C) Accuracy of time multiplexed incoherent accumulation. We accumulate four different 1 ns pulses as sketched in (A) with a delay of 1.86 ns. The measured signal is distributed around the calculated signal with a standard deviation of 1.1% which is comparable the noise within the measurement setup itself.

artificial neural networks, each component of x corresponds to the activation of a neuron, thus we assume $x_i \geq 0$. MVMs with arbitrary input vectors can be computed by using a reference vector component [35]. In contrast, the weights connecting the various neurons of the ANN can be positive and negative. Hence, M is an arbitrary real valued $m \times n$ matrix. The $m \times n$ matrix is divided into several $m \times n'$ matrices with $n' \ll n$. Then, the matrix vector multiplication is performed stepwise where the first sub matrix is multiplied with the first elements of the vector, analogously for the other sub matrices. The different sub results are then accumulated by the photodetector to obtain the correct result without requiring any electro-optical conversion and electrical processing. In this way, only a $m \times n'$ crossbar array is required to carry out the same processing as obtained from the significantly larger $m \times n$ matrix, greatly reducing the complexity of the photonic circuit.

We propose a design for a matrix cell of the photonic crossbar array which encodes the matrix weight into a Mach–Zehnder modulator (MZM) as shown in Figure 3B. This enables a fast modulation of the matrix elements which is crucial for the time multiplexed MVMs. Moreover, it allows for reference computation without inducing any additional loss by performing balanced detection between the two output waveguides. In this scheme, a matrix weight of zero is encoded by setting the MZM to equal splitting between Output+ and Output−. Similarly, negative weights are implemented by guiding are larger fraction of the Input to Output- and vice versa. State of the art MZM achieves modulation speeds of 100 GHz on a comparably small footprint [37–40]. Assuming the

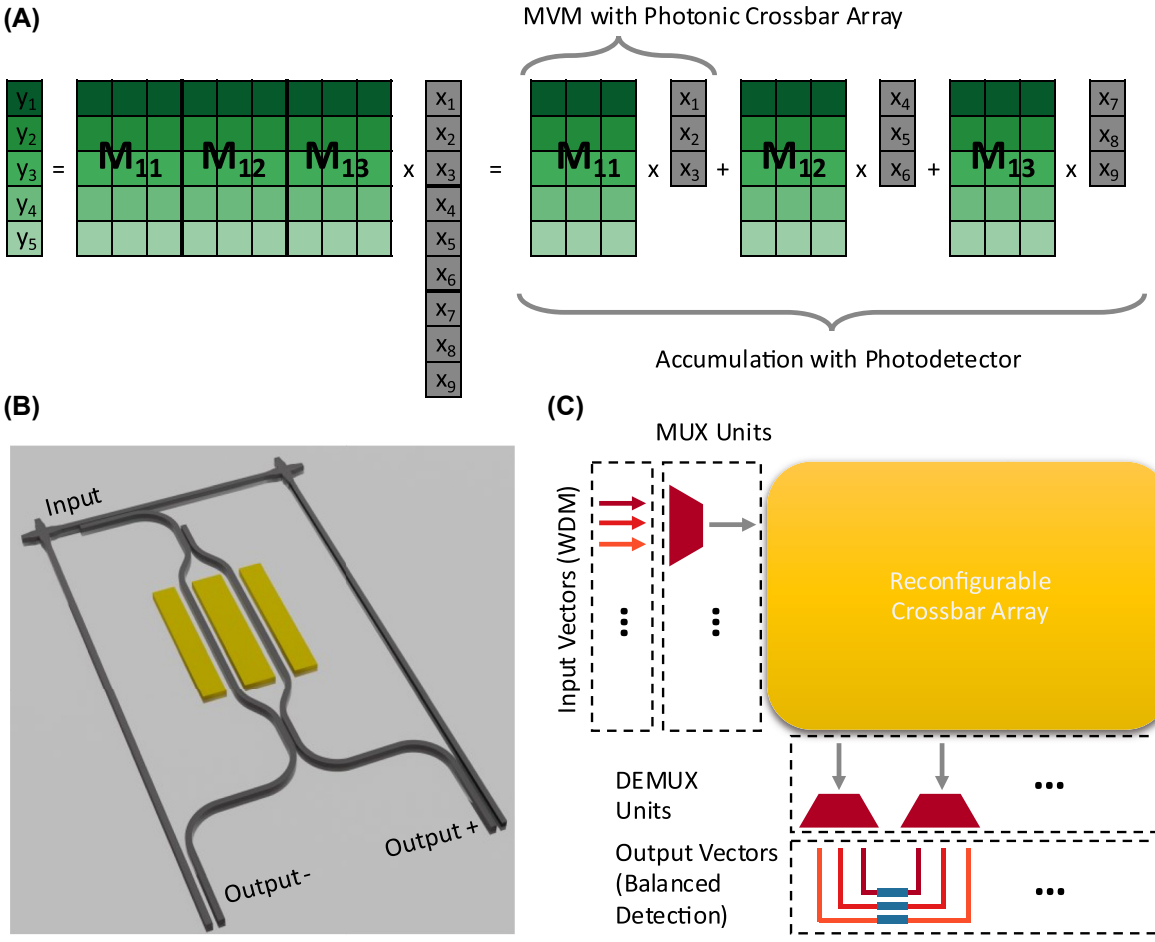

Figure 3: Large scale photonic matrix vector multiplication scheme. (A) Time multiplexed matrix vector multiplication (MVM). We decompose the large-scale matrix vector multiplication into several small MVMs which the photonic crossbar array computes. We sum the intermediate results by charge accumulation inside the photodetector. In this way, no additional electrical processing is required. (B) Concept of an active, arbitrary valued matrix cell. We design matrix cells deploying an MZM as the matrix weight. In this way, the weight can be modulated as fast as the input vector, which is crucial for this matrix architecture. Moreover, the MZM allows for lossless reference computation. (C) Sketch of the complete matrix processor. The processor computes several MVMs in parallel by encoding each vector in a different wavelength range. Before sending the vectors to the crossbar array, they are multiplexed together and afterwards demultiplexed again to obtain the individual output vectors. Finally, balanced detection is performed to obtain the correct result.

silicon-organic hybrid MZM in [40], the crossbar cell size would be in the order of 0.05 mm^2.

Figure 3C sketches the photonic circuit of the complete system consisting of MZMs to modulate the input vectors, a quickly reconfigurable photonic crossbar array, comparably slow photodetectors for temporal accumulation and wavelength multiplexer. Wavelength division multiplexing enables several parallel computational channels which further increases the speed of the photonic matrix processor [35]. We estimate the ultimate performance capabilities of the time multiplexed matrix processor architecture based on the characterized 10 MHz photodetector, 100 GHz MZMs and a 16 × 64 crossbar array. The photodetector precisely accumulates the optical pulses within a time of 10 ns, corresponding to 1000 pulses due to the modulation speed of the MZMs. In this way the size of the crossbar array is virtually increased to 16,000 × 64 but only requires 1024 MZMs instead of more than a million as its equivalent physical counterpart. Moreover, it greatly reduces the overall loss of the system in comparison to a physical, incoherent PCA of the same size, since the splitter induced loss scales as $1/n$ where n is the number of physical input waveguides to the PCA. Even though the time-multiplexing decreases the theoretical maximal speed of the system (due to the need for slower than state-of-the-art photodetectors), wavelength division multiplexing still unlocks exceptional computational power. Frequency combs can generate the required input wavelengths over a wavelength range of

1500–1650 nm [9]. Assuming a 100 GHz spacing between the wavelengths of one channel to avoid interference, each optical matrix vector multiplication requires 16 (#inputs) x 0.8 nm (#channel spacing) = 12.8 nm optical bandwidth, so 11 vectors can be computed in parallel. The matrix processor performs MVMs with 16,000 × 64 matrices at a speed of 10 MHz on an estimated crossbar array area of 51.2 mm^2, leading to 112.64 trillion multiply and accumulate operations per second. The main advantage of this approach is that the matrix vector multiplication is performed fully optically without any intermediate electrical processing steps. A physical implementation would thus greatly reduce the bottleneck of optical data processing caused by electro optical conversion.

5 Summary

Photonic computing is a promising approach to fulfil the ever-growing demand on computational performance arising from the use of artificial neural networks. The combination of wavelength division multiplexing and in-memory computing enables matrix-vectors-multiplications at unprecedented computation speeds and low latency times [7, 9, 41]. However, scaling the photonic circuit to perform large-scale MVMs remains challenging, due to the physical size of the photonic components and fabrication imperfections. We propose a novel computation scheme for photonic matrix processors, which allows one to virtually increase the size of the MVMs without suffering from the drawbacks of large photonic circuits. The computation scheme uses time multiplexed incoherent accumulation, which encodes the pulses in both the frequency and time domain. We characterize this scheme with a 10 MHz detector and 1 ns pulses, allowing integration in the time domain up to 10 ns with an accuracy of 98.9%. Assuming a 100 GHz MZM [40], a 1500–1650 nm frequency comb [9] and a reasonable sized physical 16 × 64 crossbar [11], the system could perform MVMs with 16,000 × 64 matrices all optically at a speed above 110 trillion multiply and accumulate operations per second.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: We acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG) through CRC 1459, the European Commission via grants 101017237, 101046878, 724707 and 899598, and the BMBV via grant HYPHONE.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

[1] M. A. A. Babiker, M. A. O. Elawad, and A. H. M. Ahmed, "Convolutional neural network for a self-driving car in a virtual environment," in *Proc. Int. Conf. Comput. Control. Electr. Electron. Eng. 2019, ICCCEEE 2019*, 2019.

[2] T. Brown, B. Mann, N. Ryder, et al., "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[3] H. Gehring, M. Blaicher, W. Hartmann, et al., "Low-loss fiber-to-chip couplers with ultrawide optical bandwidth," *APL Photonics*, vol. 4, no. 1, p. 010801 1–010801 7, 2019.

[4] Z. Lu, H. Yun, Y. Wang, et al., "Broadband silicon photonic directional coupler using asymmetric-waveguide based phase control," *Opt. Express*, vol. 23, no. 3, p. 3795, 2015.

[5] H. Yang, P. Zheng, G. Hu, R. Zhang, B. Yun, and Y. Cui, "A broadband, low-crosstalk and low polarization dependent silicon nitride waveguide crossing based on the multimode-interference," *Opt. Commun.*, vol. 450, pp. 28–33, 2019.

[6] S. Siontas, H. Wang, D. Li, A. Zaslavsky, and D. Pacifici, "Broadband visible-to-telecom wavelength germanium quantum dot photodetectors," *Appl. Phys. Lett.*, vol. 113, no. 18, pp. 181101-1–181101-4, 2018.

[7] A. Van Laer, M. R. Madarbux, P. M. Watts, and T. M. Jones, "Exploiting silicon photonics for energy-efficient heterogeneous parallel architectures (SiPhotonics'2014)". https://www.cl.cam.ac.uk/~tmj32/papers/docs/vanlaer14-siphotonics.pdf.

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

[9] J. Feldmann, N. Youngblood, M. Karpov, et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.

[10] A. N. Tait, T. Ferreira de Lima, E. Zhou, et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017.

[11] C. Ramey, "Silicon photonics for artificial intelligence acceleration: HotChips 32," in *2020 IEEE Hot Chips 32 Symp. HCS 2020*, 2020.

[12] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.

[13] S. Bandyopadhyay, A. Sludds, S. Krastanov, et al., *Single Chip Photonic Deep Neural Network with Accelerated Training*, 2022, pp. 1–21 [Online]. Available at: http://arxiv.org/abs/2208.01623.

[14] G. Dabos, D. V. Bellas, R. Stabile, et al., "Neuromorphic photonic technologies and architectures: scaling opportunities and performance frontiers [Invited]," *Opt. Mater. Express*, vol. 12, no. 6, p. 2343, 2022.

[15] H. Zhou, J. Dong, J. Cheng, et al., "Photonic matrix multiplication lights up photonic accelerator and beyond," *Light Sci. Appl.*, vol. 11, no. 1, 2022. https://doi.org/10.1038/s41377-022-00717-8.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016, 2016, pp. 770–778.
[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.
[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int. Conf. Learn. Represent. ICLR 2015 – Conf. Track Proc.*, 2015, pp. 1–14.
[19] A. Krizhevsky, I. Sutskever, and G. E. H. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2012, pp. 1097–1105.
[20] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017, pp. 5999–6009, 2017.
[21] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 – 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. – Proc. Conf.*, vol. 1, 2019, pp. 4171–4186.
[22] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=dOad5HoAAAAJ&citation_for_view=dOad5HoAAAAJ:W7OEmFMy1HYC.
[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners." Technical report, OpenAi, 2019/2/14.
[24] J. J. Hopfield and D. W. Tank, "'Neural' computation of decisions in optimization problems," *Biol. Cybern.*, vol. 52, no. 3, pp. 141–152, 1985.
[25] M. Prabhu, C. Roques-Carmes, Y. Shen, et al., "Accelerating recurrent Ising machines in photonic integrated circuits," *Optica*, vol. 7, no. 5, p. 551, 2020.
[26] E. H. L. Aarts and J. H. M. Korst, "Boltzmann machines for travelling salesman problems," *Eur. J. Oper. Res.*, vol. 39, pp. 79–95, 1989.
[27] C. Roques-Carmes, Y. Shen, C. Zanoci, et al., "Heuristic recurrent algorithms for photonic Ising machines," *Nat. Commun.*, vol. 11, no. 1, pp. 1–8, 2020.
[28] N. A. Aadit, A. Grimaldi, M. Carpentieri, et al., "Massively parallel probabilistic computing with sparse Ising machines," *Nat. Electron.*, vol. 5, no. 7, pp. 460–468, 2022.
[29] T. Honjo, T. Sonobe, K. Inaba, et al., "100, 000-spin coherent Ising machine," *Sci. Adv.*, vol. 7, no. 40, pp. 1–8, 2021.
[30] K. Tatsumura, M. Yamasaki, and H. Goto, "Scaling out Ising machines using a multi-chip architecture for simulated bifurcation," *Nat. Electron.*, vol. 4, no. 3, pp. 208–217, 2021.
[31] M. Yamaoka, C. Yoshimura, M. Hayashi, T. Okuyama, H. Aoki, and H. Mizuno, "A 20k-spin Ising chip to solve combinatorial optimization problems with CMOS annealing," *IEEE J. Solid State Circ.*, vol. 51, no. 1, pp. 303–309, 2016.
[32] C. Ríos, N. Youngblood, Z. Cheng, et al., "In-memory computing on a photonic platform," *Sci. Adv.*, vol. 5, no. 2, pp. 1–10, 2019.
[33] 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.*, vol. 26, no. 1, pp. 1–11, 2020.
[34] F. Brückerhoff-Plückelmann, J. Feldmann, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "Chalcogenide phase-change devices for neuromorphic photonic computing," *J. Appl. Phys.*, vol. 129, no. 15, pp. 1–9, 2021.
[35] F. Brückerhoff-Plückelmann, J. Feldmann, H. Gehring, et al., "Broadband photonic tensor core with integrated ultra-low crosstalk wavelength multiplexers," *Nanophotonics*, vol. 11, no. 17, pp. 1–10, 2022.
[36] R. Hamerly, A. Sludds, S. Bandyopadhyay, et al., "Netcast: low-power edge computing with WDM-defined optical neural networks," *arXiv*, vol. 14, no. 8, pp. 1–11, 2022.
[37] C. Wang, M. Zhang, X. Chen, et al., "Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages," *Nature*, vol. 562, no. 7725, pp. 101–104, 2018.
[38] L. Alloatti, R. Palmer, S. Diebold, et al., "100 GHz silicon-organic hybrid modulator," *Light Sci. Appl.*, vol. 3, pp. 5–8, 2014.
[39] Y. Gui, B. M. Nouri, M. Miscuglio, et al., "100 GHz micrometer-compact broadband monolithic ITO Mach – Zehnder interferometer modulator enabling 3500 times higher packing density," *Nanophotonics*, vol. 11, no. 17, pp. 4001–4009, 2022.
[40] C. Kieninger, C. Füllner, H. Zwickel, et al., "SOH Mach-Zehnder modulators for 100 GBD PAM4 signaling with sub-1 dB phase-shifter loss," in *Optical Fiber Communication Conference (OFC) 2020*, pp. 1–3, 2020.
[41] X. Xu, M. Tan, B. Corcoran, et al., "11 TOPS photonic convolutional accelerator for optical neural networks," *Nature*, vol. 589, no. 7840, pp. 44–51, 2021.