Towards a Photonic Integrated Linear Algebra Processor for Complex Matrix Multiplication and Inversion

Qixiang Cheng,[1] Minjia Chen,[1] Masafumi Ayata,[2] Mark Holm,[2] Richard Penty[1]
[1]*Electrical Engineering Division, Department of Engineering, University of Cambridge*
[2]*Radio Basestation Systems Department, Huawei Technologies (Sweden) AB, Gothenburg, Sweden*
qc223@cam.ac.uk

Abstract: We propose the first on-chip optical linear algebra processor for complex-valued matrix multiplication and inversion, incorporating III-V gain blocks. Architecture properties and design considerations are discussed, with numerical analyses showing its superiority over electronic processors.

1. Introduction

Complex matrix multiplication and inversion are two fundamental but computationally expensive linear algebra operations that are vital for use in complex-valued neural networks [1], and wireless communications [2]. Photonic integration has been proven as a powerful platform for enabling optical linear algebra computing in recent years, creating a new framework for information processing machines. Albeit that a number of advanced photonic integrated processors have been demonstrated for ultrafast matrix multiplication in artificial neural networks [3, 4], none of them have been shown to be capable of handling complex-valued computations. Developing complex-valued optical matrix inverters is even more challenging since both phase-sensitive designs and gain-integration are indispensable. Enabled by the proliferation of heterogeneous integration technologies, we propose the first on-chip optical linear algebra processor that allows complex-valued matrix multiplication and inversion. The proposed photonic processor is highly scalable, thanks to the semiconductor optical amplifier (SOA) add-ons. The introduction of wavelength multiplexing in a matrix-vector unit facilitates a low-power and compact full matrix-matrix multiplier, which can readily be converted to a matrix-matrix inverter following Richardson's method [5, 6]. Numerical analyses are presented, indicating the inversion rate of the proposed photonic integrated matrix inverter can be a few GHz (orders of magnitude higher than an electronic processor), with over an order of magnitude higher energy efficiency.

2. Design of the linear algebra processor

The design of the $N \times N$ linear algebra processor is illustrated by Figure 1a. It is a column-oriented matrix-vector multiplier which calculates $\boldsymbol{y} = \boldsymbol{Mx}$ by $\boldsymbol{y} = \sum_{i=1}^{N} x_i \boldsymbol{M^{(i)}}$, where $\boldsymbol{M^{(i)}}$ is the i^{th} column of the matrix. The summation is implemented by cascaded 4-port couplers (termed as 2×2 adders). The multiplication is done by imprinting the complex-valued matrix weights on the input vectors via Mach-Zehnder Interferometers (MZIs). The amplitude of the weights is determined by setting the differential phase shift between the two MZI arms, while the phase is determined by adding an extra identical biased phase shift on both arms. Signal fan-out is realized by employing cascaded 3-port 3dB couplers. However, inherent and excess losses of the chip will lead to limitations and deviations of the computation. We thus use SOA blocks for on-chip amplification. A number of heterogeneous integration schemes can be leveraged [7, 8], such as flip-chip bonding, micro transfer-printing, and wafer bonding. Wavelength multiplexing is proposed to implement a full matrix-matrix processor in one matrix-vector unit, where the 1^{st} to the n^{th} columns of input matrix $\boldsymbol{X}$ are represented by signals centred at $\lambda_1 - \lambda_n$, realizing a low-power and compact configuration.

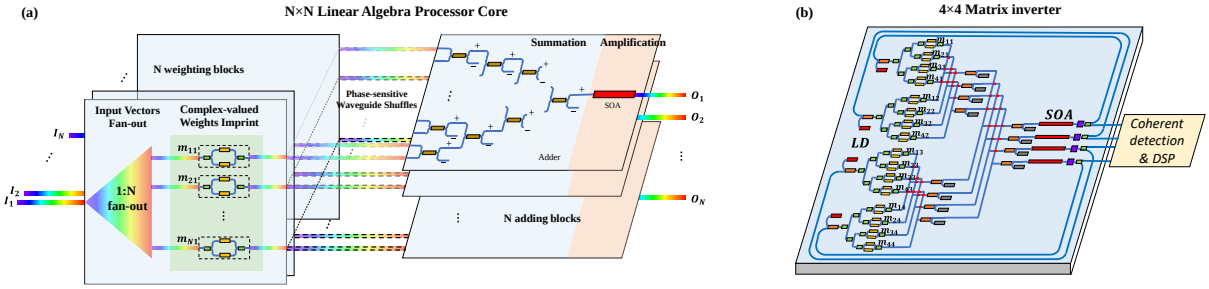

Figure 1 (a) Proposed design of the photonic linear algebra processor. (b) Conceptual figure of an integrated 4×4 inverter, in which Laser diodes (LD), SOAs, and detectors can be monolithically integrated on-chip.

A matrix inverter is implemented following Richardson's method, as illustrated by Figure 1b. Calculating the inverse of $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ is equivalent to solving $\boldsymbol{AX} = \boldsymbol{I_n}$. Richardson's method ($\omega = 1$) is:

$$\boldsymbol{X^{(k+1)}} = (\boldsymbol{I_n} - \omega \boldsymbol{A})\boldsymbol{X^{(k)}} + \omega \boldsymbol{I_n}, \tag{1}$$

where $\boldsymbol{X^{(0)}} = \boldsymbol{0}$. Since $\boldsymbol{I_n} - \boldsymbol{A}$ is a fixed matrix in each iteration, Eq.1 can be simplified to:

$$\boldsymbol{X}^{(k+1)} = \boldsymbol{M}\boldsymbol{X}^{(k)} + \boldsymbol{I}_n, \tag{2}$$

where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{A}$. Convergence requires $\boldsymbol{A}$ to be positive-definite. The initial input is identity matrix $\boldsymbol{I}_n$ whose columns are continuous-wave signals centred at $\lambda_1 - \lambda_n$. $\boldsymbol{M}\boldsymbol{X}^{(k)}$ is a matrix-matrix multiplication which is operated by the processor core, sent back to the processor inputs via phase-sensitive loops, and added to the input identity matrix. It's worth noting that the physical length of the looped-back line determines the processing rate of the matrix inverter. Thus, the monolithic integration technology can yield super-fast photonic processors by having on-chip millimetre-range waveguide loops. Once converged, the inverted matrix is output using coherent detectors. Figure 2a shows an example of the convergence of a 2×2 matrix after the first iteration and the 50th iteration (set converged condition).

3. Numerical analyses

Three critical figures of merits for linear algebra processors are accuracy, speed and power efficiency. The accuracy is defined as 1-ε, where ε is the relative error defined as $\varepsilon = \|\boldsymbol{Y}_{sim} - \boldsymbol{Y}_{ideal}\|/\|\boldsymbol{Y}_{ideal}\| \times 100\%$ ($\boldsymbol{Y}_{sim}$ is the simulated processing output, $\boldsymbol{Y}_{ideal}$ is the theoretical output, and $\|\cdot\|$ is the 2-norm of a matrix). Main error sources include (1) Quantization error that is due to the finite resolution of digital-to-analogue converters (DACs), which deviates the imprinted matrix weights. This can be compensated by using high-resolution DACs. (2) ASE noise from the SOAs, which is a white noise added to the signals from each amplification loop. Bandpass filters (BPFs) are used to suppress the ASE noise power. (3) Shot noise and thermal noise from the coherent detection. As a result of the lossless operation of the processor, thermal noise is negligible. (4) Finite iterations in matrix inversion computation. This can be resolved by setting a minimum number of iterations to ensure the convergence condition is satisfied. With an analytical model build-up, we show in Fig. 2b the computation accuracy of a 32×32 photonic processor (both as a multiplier and inverter) can exceed 98%, together with contributions of different error sources. It can be seen that ASE noise is the dominating factor. Clearly, this can be ameliorated by further reducing the passband of the BPF. We also investigate the inversion accuracy as a function of the processor size and filter passband, as shown in Fig. 2c, indicating that over 90% accuracy can be achieved for a 256×256 optical matrix inverter as long as the on-chip bandpass filters are properly designed.

With a 10mm-level on-chip waveguide loop, the photonic processor can achieve an inversion rate of ~2GHz, which is orders of magnitude faster than an electronic processor (a complex-valued 4×4 electronic processor operates at a few hundred MHz [9]). The power efficiency (pJ/FLOP) of a 32×32 photonic inverter is estimated to be around 15pJ/FLOP, over 10 times higher than that of an electronic processor [10], showing great superiority of the photonic matrix inverters over traditional electronic chips.

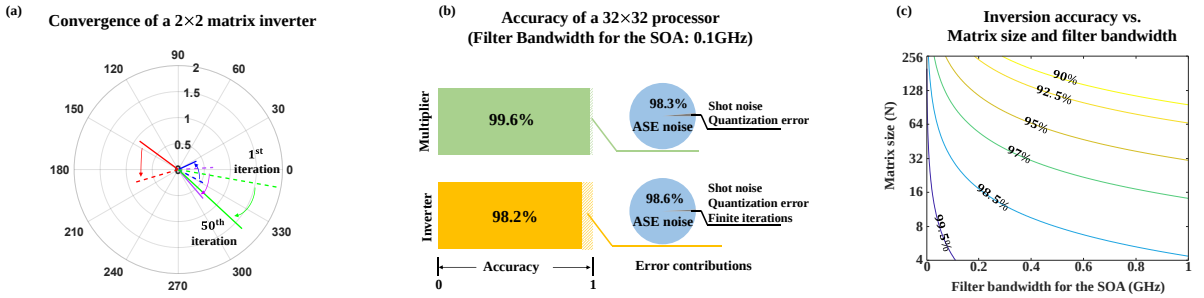

Figure 2 (a) Example of the convergence of a 2×2 matrix. (b) Accuracy and error contributions for a 32×32 matrix multiplier and inverter. (c) Inversion accuracy as a function of the processor size and filter passband.

4. Conclusions

The first $N \times N$ photonic integrated linear algebra processor suitable for complex-valued matrix multiplication and inversion is proposed. Numerical analyses show significant speed and power efficiency improvement over traditional electronic processors.

References

[1] J. Bassey *et al.*, "A Survey of Complex-Valued Neural Networks," *arXiv:2101.12249 [cs, stat]*, Jan. 2021, Accessed: Aug. 16, 2021.
[2] P. Yang *et al.*, "6G Wireless Communications: Vision and Potential Techniques," *IEEE Network*, vol. 33, no. 4, pp. 70–75, Jul. 2019.
[3] Q. Cheng *et al.*, "Silicon Photonics Codesign for Deep Learning," *Proceedings of the IEEE*, vol. 108, no. 8, pp. 1261–1282, Aug. 2020.
[4] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photon*, vol. 11, no. 7, pp. 441–446, Jul. 2017.
[5] H. Rajbenbach *et al.*, "Optical implementation of an iterative algorithm for matrix inversion," *Appl. Opt.*, vol. 26, no. 6, p. 1024, Mar. 1987.
[6] D. S. Watkins, *Fundamentals of matrix computations*, 2nd ed. New York: Wiley-Interscience, 2002.
[7] T. Matsumoto *et al.*, "Hybrid-Integration of SOA on Silicon Photonics Platform Based on Flip-Chip Bonding," *Journal of Lightwave Technology*, vol. 37, no. 2, pp. 307–313, Jan. 2019.
[8] B. Haq *et al.*, "Micro-Transfer-Printed III-V-on-Silicon C-Band Semiconductor Optical Amplifiers," *Laser & Photonics Reviews*, vol. 14, no. 7, p. 1900364, Jul. 2020.
[9] L. Sun *et al.*, "Design and VLSI Implementation of a Reduced-Complexity Sorted QR Decomposition for High-Speed MIMO Systems," *Electronics*, vol. 9, no. 10, p. 1657, Oct. 2020.
[10] M. Salmani *et al.*, "Photonic computing to accelerate data processing in wireless communications," *Opt. Express, OE*, vol. 29, no. 14, pp. 22299–22314, Jul. 2021.