Chip-scale optical matrix computations for PageRank algorithm

Hailong Zhou, Yuhe Zhao, Gaoxiang, Xu, Xu Wang, Zhipeng Tan, Jianji Dong and Xinliang Zhang

***Abstract*—Matrix computations are indispensable tools in science and engineering, while the electronic matrix computations suffered from limited speed. Alternatively, the optical methods offered a high-speed solution. The optical matrix computation is impractical unless it is capable of extending to a large scale, reconfiguring for a general purpose and operating with high efficiency. Here, we report and experimentally demonstrate an optical matrix computing processor based on an integrated linear optical network. The proposed photonic processor is capable of performing fundamental matrix computations including *XB=C*, *AB=X* and *AX=C*, where *A*, *B*, *C* are known matrices, and *X* is the matrix to be solved. An optical PageRank algorithm is further demonstrated based on the matrix computing processor for the first time. Our demonstration offers an optical method to achieve matrix computations and PageRank algorithm. Meanwhile, it suggests great potential for chip-scale fully programmable matrix computations with self-configuring methods.**

***Index Terms*—Optical computing, silicon photonics, gradient methods**

I. Introduction

MATRIX computations form the most widely used computational tools in science and engineering. Matrix computations, including the solution of systems of linear equations, least squares problems, and algebraic eigenvalue problems, govern the performance of many applications on vector and parallel computers [1-3]. The nonnegative matrix, as a special kind of matrix whose elements are all nonnegative, is widely applied in image recognition [4], voice recognition [5], biomedical and chemical sciences [6]. One of the typical applications of nonnegative matrices is the PageRank algorithm [7-9]. The PageRank algorithm, used in the Google search engine, greatly improves the results of web search by taking into account the link structure of the web. Traditionally, these matrix computations were solved by electronic digital signal processing, of which the speed is limited by the electronic clock rate. A solution to achieve high-speed matrix computations is to implement these functions directly in an optical network [10-18], although not mature yet. The optical matrix computations show distinct advantages on computing speed and power consumption compared to electronic matrix computations. For examples, discrete Fourier transforms and vector multiplier were performed at the speed of light using a fully parallel, high-speed incoherent optical method [10, 11], while it is difficult to update the matrix at high frame rates with currently available two-dimensional spatial light modulators. Optical array modulators, such as electrooptic modulations, direct driven LED array, or acoustooptic Bragg cells, were used to perform matrix multiplications with faster frame rates [12-14]. The matrix-matrix multiplication could be optionally accomplished using a photorefractive crystal [15-17] and nonlinear material [18]. Although the photonic implementations of these functions have been reported, they are bulky and complex with bench-top optics, thus it is difficult to expand the network and difficult to reconfigure the flexible computation functions. For large-scale and general-purpose matrix computations, an optical matrix computing processor should be able to perform multiple functions efficiently with high reconfigurability and expandability.

In this paper, we report and experimentally demonstrate an optical matrix computing processor based on gradient descent algorithm, which can perform fundamental matrix computations including *XB=C*, *AB=X* and *AX=C*, where *A*, *B*, *C* are known matrices, and *X* is the matrix to be solved. The functions are programmable by self-configuring method. The training is accomplished using the numerical gradient descent algorithm modified from deep learning. The application for optical PageRank algorithm based on the chip is further demonstrated for the first time. Our demonstration offers an optical method to implement matrix computations and PageRank algorithm. It suggests great potential for chip-scale fully programmable optical matrix computations with self-configuring methods.

Manuscript received xxxx, xx, xxxx; revised xxxx xxxx , xxxx ; accepted xxxx xxxx, xxxx. Date of publication xxxx xxxx; date of current version xxxx, xxxx. This work was supported in part by National Key Research and Development Project of China (2018YFB2201901), in part by National Natural Science Foundation of China (61622502, 61805090), in part by China Postdoctoral Science Foundation(2017M622419), in part by State Key Laboratory of Advanced Optical Communication Systems and Networks, Shanghai Jiao Tong University, China (2019GZKF03002). H. Zhou and Y. Zhao contributed equally to this paper. (Corresponding authors: Zhipeng Tan, Jianji Dong)

H. Zhou, Y. Zhao, G. Xu, X. Wang, Z. Tan, J. Dong and X. Zhang is with the Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: tanzhipeng@hust.edu.cn; jjdong@hust.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2943347, IEEE Journal of Selected Topics in Quantum Electronics

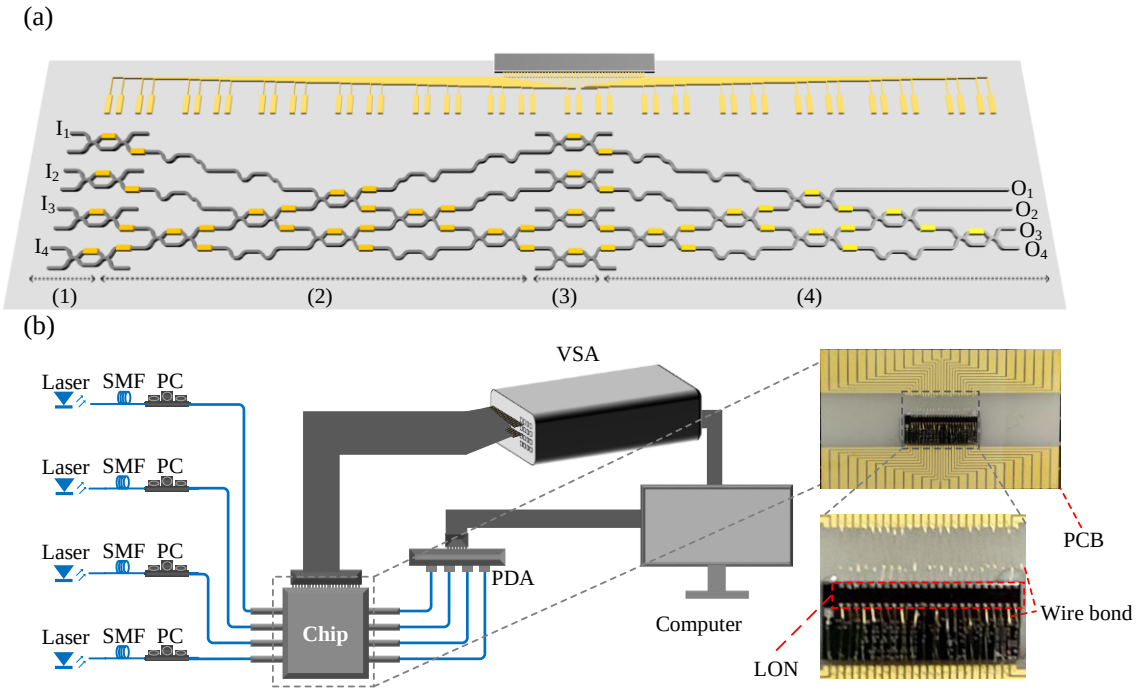

Fig. 1. Optical matrix computing processor. (a) The detailed structure of LON chip. The LON chip contains 48 phase shifters and 20 Mach-Zehnder interferometers (MZIs) in total. It can be divided to four functional parts, which carry out the following: (1) tuning the input power; Parts (2), (3) and (4), as a whole, are used to implement an arbitrary matrix transformation based on singular value decomposition. (b) Schematic of the experimental set-up. Four tunable continuous wave lasers with the same wavelength at around 1550 nm are launched into our LON chip through a V-groove fibre array (VGA). The polarizations of input/output light are optimized by in-line polarization controllers (PCs). Photons emerging from the device are collected by another VGA and a 4-channel photodetector array (PDA). The LON chip contains 48 phase shifters which are controlled by a voltage source array (VSA). The devices of VSA and PDA are connected and controlled by the same computer. The inset shows the details of chip packaging. The SOI chip is glued on a ceramic-wafer printed circuit and the pads are wire-bonded to the printed circuit board (PCB).

II. Principle

A. Experimental set-up

The matrix computing processor, structured with a linear optical network (LON) shown in Fig. 1(a) [19], is fabricated on the commercial silicon-on-insulator (SOI) wafer (APPENDIX A). The LON chip consists of four parts. Part (1) is constitutive of four Mach-Zehnder interferometers (MZIs), which are used to tune the input power. Parts (2) and (4) can both perform an arbitrary unitary matrix transformation [19-26]. Other types of structured meshes, such as hexagonal and quadrilateral cell structures, can alternatively perform the unitary matrix transformation [27-29]. Part (3) can perform an arbitrary diagonal matrix transformation. Parts (2), (3) and (4), as a whole, are used to implement an arbitrary matrix transformation based on singular value decomposition [30]. The LON chip containing 48 phase shifters and 20 MZIs totally. The experimental set-up is depicted in Fig. 1(b). Four independent laser beams at around 1550 nm are launched into our LON chip through a V-groove fibre array (VGA). The polarizations of input/output light are optimized by in-line polarization controllers (PCs). Another VGA and a 4-channel photodetector array (PDA) are used to receive the output light from the chip. All the phase shifters in the LON chip are driven by a voltage source array (VSA). The PDA and voltage sources are connected and controlled by the same computer. In our design, the four laser beams without phase-locked loop are incoherent, that means the loaded transmission matrix of the LON chip is applied on the power of light and can be an arbitrary nonnegative matrix (APPENDIX B). If coherent beams are used, the transmission matrix is applied on the complex amplitude of light and can be arbitrary. Based on the set-up, multiple fundamental matrix computations can be implemented by self-configuration. According to the different functions, a suitable cost function (CF) should be first defined for a successful training. Then the only training target is to make the defined CF maximum using the numerical gradient descent algorithm modified from deep learning [30, 31]. In the following, the LON chip is reconfigured to achieve three different functions. And a further application of optical PageRank algorithm based on the chip is demonstrated for the first time.

B. Data reading and writing

The output vector of the LON chip can be directly read out from the output powers. And the input vector can be read out and written in according to the applied voltages on the phase shifters of Part (1) (APPENDIX C). The column vectors of loaded matrix (X_{ONN}) can be simply read out from the output ports by inputting standard unit vectors of e_1, e_2, …, e_n in sequence (namely, inputting the identity matrix). Similarly, a known transmission matrix of A can be written into the LON chip by training the network to make $XI = A$, where I is the identity matrix. The principle of training is presented below (Matrix computation of XB=C).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2943347, IEEE Journal of Selected Topics in Quantum Electronics

C. *Matrix computation of XB=C*

Matrix computation of $XB=C$ is an issue to find a solution for X when B and C are given. Here, X is a $n\times n$ dimensional matrix, B and C are $n\times m$ dimensional matrices. We rewrite the equation in the form of column vectors by

$$X\left[B_1 \quad B_1 \quad \cdots \quad B_m\right]=\left[C_1 \quad C_1 \quad \cdots \quad C_m\right]. \tag{1}$$

To find the solution with our chip, we tune the transmission matrix of the chip (X_{ONN}) to make $X_{ONN}B=C$ with a self-configuring algorithm (APPENDIX C). B_1 , B_2 , …, B_m are loaded into the chip in sequence by tuning the phase shifters in Part (1) and the corresponding output power distributions are measured, recorded by $C_{\exp}=\left[C_{\exp 1} \quad C_{\exp 2} \quad \cdots \quad C_{\exp m}\right]$. Obviously, the loaded matrix of X_{ONN} will be equal to X when $C_{\exp}=C$. In order to train the network to reach the target of $C_{\exp}=C$, we define the CF by

$$\mathrm{CF}=\frac{\left|C\bullet C_{\exp}\right|}{\|C\|\left\|C_{\exp}\right\|}. \tag{2}$$

Here, the operation ' $\bullet$ ' means the scalar product of two vectors or matrices, and ' $\|\cdot\|$ ' is the Frobenius norm of a vector or a matrix. The CF is ranged from 0 to 1, where CF=1 or 0 represents that the experimental and theoretical results are consistent or irrelevant. A numerical gradient descent algorithm is employed to train the network. After the LON is well trained, the solution of $X=X_{ONN}$ can be directly read out from the chip.

The method can be also used to write a known transmission matrix of A into the LON chip by training the network to make $XI=A$. Once the targeted transmission matrix is loaded, other types of matrix computations, such as $AB=X$, $AX=C$ and the application of PageRank algorithm, can be performed with our chip. The principle is described in the following.

D. *Matrix computation of AB=X*

To solve the equation of $AB=X$, the first step is to write the matrix of A into the LON chip, which can be accomplished by training the equation of $XI=A$. After loading the matrix of A, the column vectors of B are loaded into the chip one by one, then the solution of X can be read out directly from the output.

E. *Matrix computation of AX=C*

To solve the equation, the first step is to load the matrix of A into the LON chip. And then we try to tune the input power distribution to make the output vectors ($C_{\exp 1}$, $C_{\exp 2}$, …, $C_{\exp m}$) parallel to C_1 , C_2 , …, C_m one by one. The CF is defined independently for each column vector, given by

$$\mathrm{CF}_i=\frac{\left|C_i\bullet C_{\exp i}\right|}{\|C_i\|\left\|C_{\exp i}\right\|}. \tag{3}$$

After the LON training is complete, X_1 , X_2 , …, X_m can be directly read out from the input power distributions and proportionally revised according to the module of output vectors.

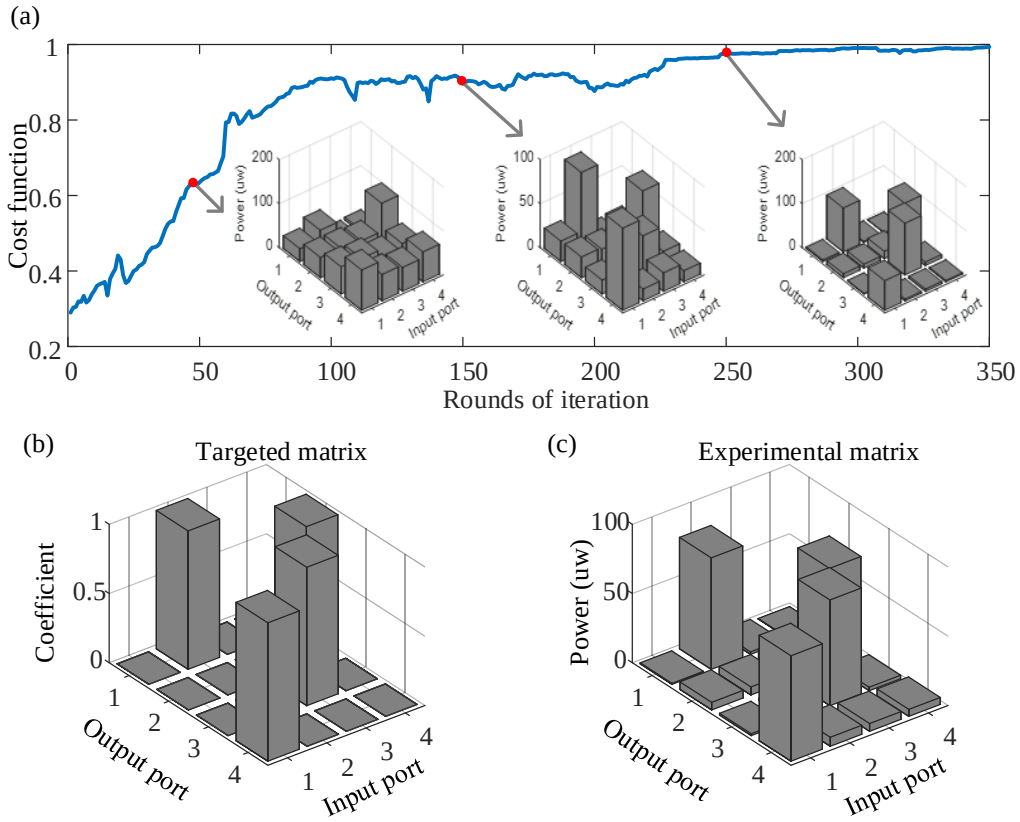

Fig. 2. Experimental results for Matrix computation of $XB=C$. (a) Cost function (CF) dependent on the rounds of iteration in training process. The insets figures show the light power distributions when the round of iteration equals 50, 150 and 250 respectively. (b) The targeted matrix. (c) The final experimental matrix.

F. PageRank algorithm.

PageRank is the first and meanwhile the best-known algorithm used by Google Search to rank web pages in their search engine results. PageRank, named after Larry Page [7-9], is a way to measure the importance of website pages by solving the equation (APPENDIX D)

$$GP = P. \tag{4}$$

where G is the Google matrix recording the links of all webpages and P is the importance of all pages. The elements of each column of Google matrix sum up to 1, so the matrix is a stochastic matrix (also called Markov matrix) [32]. It is a variant of the eigenvector centrality measure used commonly in network analysis. Traditionally, PageRank can be computed either iteratively or algebraically. The iterative method can be viewed as the power iteration method by

$$P(n+1) = GP(n). \tag{5}$$

where n and $n+1$ mean the present round of iteration and the next round of iteration respectively. Here, we use an optical method to finish the iteration algorithm. The Google matrix can be loaded into our LON chip by data writing. And then the PageRank scores can be obtained by iteratively updating the input vector with the output vector until problem converges. The CF is defined by

$$\mathrm{CF} = \frac{\left|P(n+1) \bullet P(n)\right|}{\left\|P(n+1)\right\|\left\|P(n)\right\|}. \tag{6}$$

After the iteration is completed, the PageRank scores can be directly read out both from the input and output.

III. 3 RESULTS

A. Matrix computation of XB=C

We first solve the equation of *XB=C* and try to write data of transmission matrix. To find the solution for this matrix equation, the LON chip is trained and optimized using numerical gradient descent algorithm, aiming to make the output matrix of $C_{\exp}$ (experimental value) close to C (expected value) as much as possible. In the experiment, we use the LON chip to solve the following equation of *XB=C*. As an example, we define *B* and C as follows

$$B = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix}, C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \tag{7}$$

This equation is applied to write a transmission matrix of C into the chip since *B* is the identity matrix. Fig. 2(a) presents the trainning process dependent on rounds of iteration. The initial loaded matrix of chip is given randomly. The output matrix of $C_{\exp}$, as shown in the insets of Fig. 2, is messy in the beginning and gradually close to the theoretical result with the trainning process. The CF, used to evaluate the trainning level, is gradually increased in the trainning process and almost reaches the theoretical maximum of 1 finally. Since $B=I$ here, the final solution of X is equal to $C_{\exp}$. Figs. 2(b) and 2(c) present the theoretical output matrix and the experimental one, which are consistent, proving the ability of solving this type of equation and loading a specified matrix with the trainning process.

B. Matrix computation of AB=X

Now we demonstrate the matrix computation of *AB=X*. The matrix of *A* is written into the LON chip by the training process shown in Fig. 2(a) and the final loaded transmission matrix is shown in Fig. 2(c). Figure 3 presents the input vectors and corresponding output vectors. The matrix equation is governed by

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0.25 \\ 0 & 1 & 1 \\ 1 & 0 & 0.75 \\ 1 & 0 & 1 \end{bmatrix} = X. \tag{8}$$

The column vectors of *B* are injected into chip from the input ports in sequence, as shown in Figs. 3(a)-3(c). Then the solution is read out directly from the power monitors of output ports. The corresponding column vectors of *X* are presented in Figs. 3(d)-3(f) respectively, consistent to the theoretical results.

C. Matrix computation of AX=C

Similarly, the transmission matrix in Fig. 2(b) is used as the initial matrix of *A*. To demonstrate this function of *AX=C*, Fig. 4 presents the targeted output vectors and corresponding results of input vectors, given by

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} X = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 0.6 \\ 1 & 1 & 0 \end{bmatrix}. \tag{9}$$

The column vectors of *X* are obtained by training the input powers to make output column vectors parallel to C_1, C_2, …, C_m one by one. For example, the input powers are optimized by tuning the phase shifters in Part (1), aiming to make the output power vector parallel to C_1. After the training of LON chip is complete, the input powers are read out and assigned to X_1. The other column vectors can be obtained in the same way. To ensure the power efficiency, the input vectors are always scaled up to make the maximum element normalized to 1. Finally all the column vectors (X_1, X_2 and X_3) are proportionally revised according to the module of output vectors. The targeted output vectors are shown in Figs. 4(a)-4(c). And the final column vectors of *X* are presented in Figs. 4(d)-4(f) respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2943347, IEEE Journal of Selected Topics in Quantum Electronics

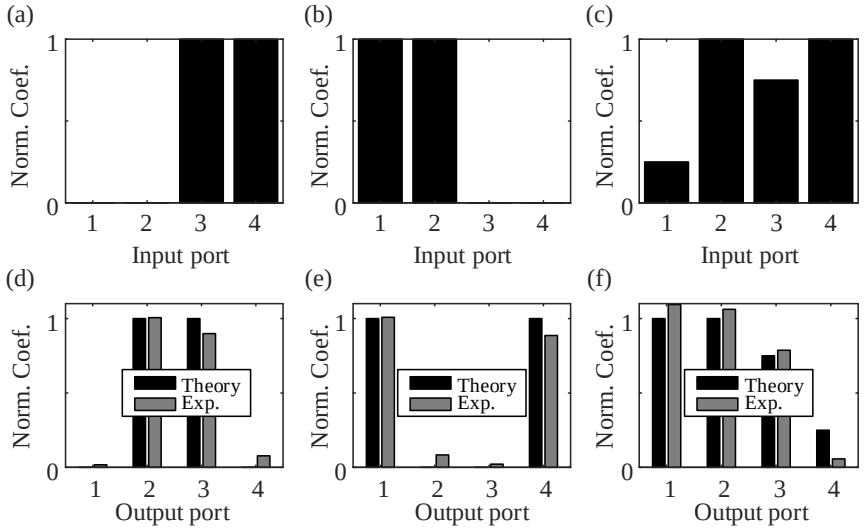

Fig. 3. Experimental results for matrix computation of $AB=X$. (a to c) Input matrix. The column vectors of B are input into the chip one by one. (d to f) Output matrix. The corresponding column vectors of solution are read out directly from the output ports.

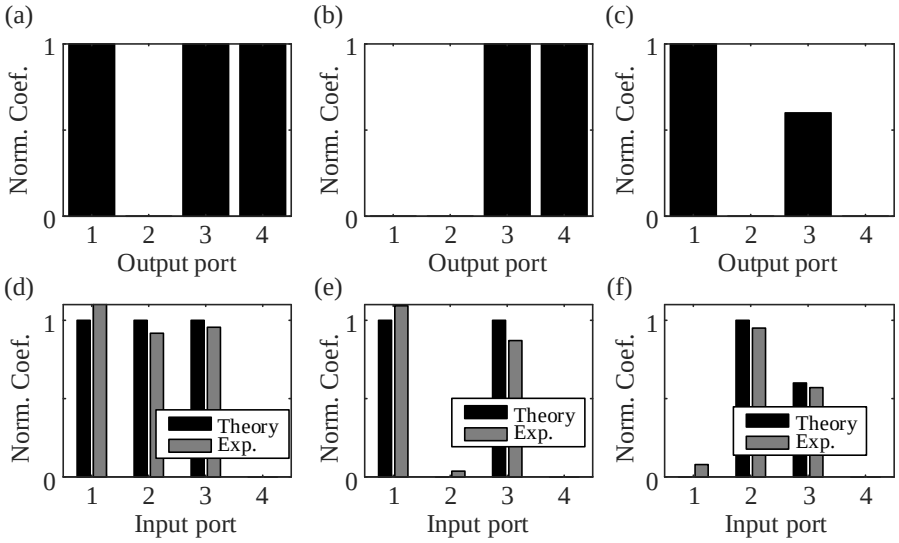

Fig. 4. Experimental results for matrix computation of $AX=C$. (a to c) Targeted output vectors. The column vectors of X are trained one by one. (d to f) Input matrix. The corresponding column vectors of solution are obtained by training the input powers.

D. PageRank algorithm.

The photonic matrix computation processor can also be used to implement PageRank algorithm with the inherent speed of light. Here, two examples are presented to demonstrate the feasibility of PageRank algorithm. In the first example, the diagram of links among four webpages is presented in Fig. 5(a), where Pages 1 and 2 have two inbound links, Page 3 has no inbound links and Page 4 has three inbound links. Intuitively, Page 4 should own the most importance and Page 3 should own the least importance. The final Google matrix can be obtained by (5) and the theoretical result is presented in Fig. 5(b) (black bars). Firstly, the Google matrix is loaded into the LON chip by solving the equation of $XI=G$ and the experimental transmission matrix is presented in Fig. 5(b) (white bars). The experimental Google matrix is consistent to the theoretical one. Then, the PageRank scores can be obtained by iteratively updating the input vector with the output vector until the problem converges. The iterative algorithm is converged quickly and the number of iterations is less than 10 in our experiment. The theoretical and experimental PageRank scores for all pages are presented in Fig. 5(c), which are consistent and match the intuitive predictions. We also demonstrate the second example of PageRank solver. The diagram of page links, Google matrix and final experimental PageRank scores are presented Figs. 5(d)-5(f) respectively. The PageRank scores are correctly solved with the LON chip compared to the theoretical predictions. PageRank for more webpages can be realized by expanding the network to a larger one.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2943347, IEEE Journal of Selected Topics in Quantum Electronics

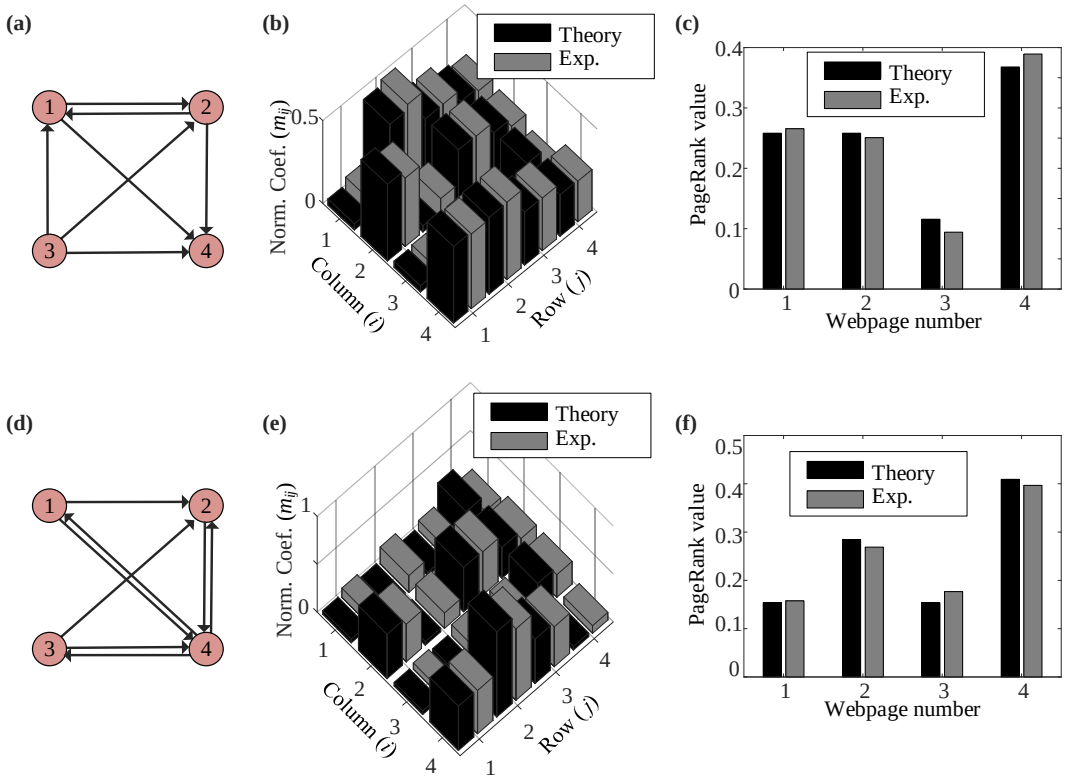

Fig. 5. Experimental results for PageRank algorithm. (a) The diagram of links among four webpages. (b) Theoretical and experimental Google matrices. (c) Final experimental PageRank scores. (d to f) The diagram of page links, Google matrix and final PageRank scores for another example.

IV. Discussion and Summary

The proposed chip-scale matrix computation processor can be reconfigured to perform matrix computations including $XB=C$, $AB=X$ and $AX=C$ with an ultra-high speed. These matrix computations are basic building blocks in linear operations. A self-configuring method is used to train the network based on numerical gradient descent algorithm. In our design, all the computations are implemented with the photonic devices and the electrical devices only offer the electronic control units. The computations are implemented with light at an ultra-high speed, so the main time cost is limited by the output time of electronic control signal provided by external drivers, the response time of photodetectors and phase shifters in the LON chip. The response speed of thermo-optic phase shifters is larger than 10 kHz [33]. And the output time of electronic control signal can be easily up to 10 kHz with high-speed power amplifiers and Digital-to-analog converters controlled by a field programmable gate array (FPGA). The training (typically less than 1000 iterations) can be completed within 100 ms. A faster training can be achieved using graphene micro-heater modulators [34] or free-carrier-depletion-based electro-optical modulators [35]. We then use this chip to demonstrate optical PageRank algorithm for the first time. The PageRank algorithm is used to calculate the importance of hundreds of millions of webpages. The PageRank algorithm or the Markov theory, also have many applications as statistical models of real-world processes, such as cruise control systems in motor vehicles, exchange rates of currencies, storage systems such as dams, and population growths of certain animal species [32]. It is extremely heavy using the electronic digital signal processor when the objects of study are massive. Here, the optical method can implement the PageRank algorithm easily with light at an ultra-high speed. And the integrated photonic network chip, which can be fabricated with the Complementary Metal Oxide Semiconductor (CMOS) process, is convenient to expand to a larger scale one, similar to the integrated electronic circuits. But the scale of network will be limited by the chip loss. It can be improved by using low-loss optical waveguides in silicon nitride integration platform [36] or optical amplifiers in InP integration platform [37]. The advantages on speed, chip size are more obvious when the matrix dimensions increase tremendously. For example, the time cost of single matrix computation in (5) for N webpages will be proportional to N^2 for electronic matrix computations. When the number of webpages increases to billions, the time cost of single matrix computation is dramatically increased, so the PageRank algorithm need completely run every a few days. In our method, the matrix computation of PageRank algorithm is performed by LON in optical domain, where the time cost for photonic matrix computations can be negligible, so the time cost is only restrained by the photodetector bandwidth (100 GHz) [38], output time of electronic control signal and response time of the phase shifters (100 GHz) [35]. The Google matrix is slow-varying with the change of webpage links. In a real situation, the new Google matrix can be loaded into the chip, iterated from the last state (old Google matrix), or the Google matrix can be real-time updated. After the Google matrix is loaded, the PageRank algorithm can be finished within 100 rounds of iteration with light.

In summary, we have designed, fabricated and demonstrated a reconfigurable optical matrix computing processor based on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2943347, IEEE Journal of Selected Topics in Quantum Electronics

an integrated LON chip, which can perform fundamental matrix computations with a self-configuring method. The training is accomplished using the numerical gradient descent algorithm modified from deep learning. The application of optical PageRank algorithm based on the chip is further demonstrated for the first time. The integrated network chip, compatible with the CMOS process, is convenient to expand to a large scale. More complex matrix computations can be accomplished using a larger network. Our demonstration offers an optical method to achieve matrix computations and PageRank algorithm with high speed, high reconfigurability and expandability. It suggests great potential for chip-scale fully programmable optical matrix computations with self-configuring methods.

APPENDIX

A. *Chip design, fabrication and characterization*

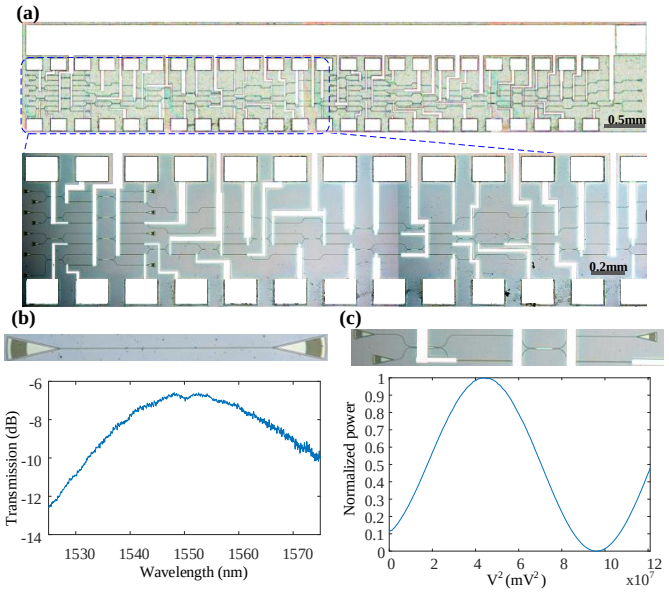

Fig. S1. Chip fabrication and characterization. (a), The micrographs of fabricated chip. (b), The transmission spectrum and the used referenced grating structure. (c), The imparted phase dependent on the applied voltage and the tested MZI structure.

The designed linear optical network (LON) chip is fabricated on the commercial SOI wafer. A passive process is employed to fabricate the structure on silicon-on-insulator (SOI) wafer with 220 nm top silicon and 2 μm thick buried oxide (BOX) substrate. The Si waveguide are 220 nm thick strip waveguide, while the grating coupler is fan-shaped with 70 nm shallow etch. Using 248 nm deep UV lithography, the structure is patterned to photoresist, followed by 70 nm partial grating etch. The remaining part of strip waveguide is then patterned and etched to BOX. Thereafter, Plasma Enhanced Chemical Vapor Deposition (PECVD) is introduced to deposit 1.2 μm pad oxide between Si waveguides and microheaters. Then a layer of 120 nm thick TiN is deposited and etched as the heater layer. Each heater unit is 160 μm long and 1 μm wide. Afterwards, Al is deposited and etched to form the metal wires and pads. The SOI chip is then glued on a ceramic-wafer printed circuit and the pads are wire-bonded to the printed circuit board. The size of chip is about 1.3 mm x 7.5 mm. The micrographs of fabricated chip are shown in Fig. S1(a). The Mach-Zehnder interferometer (MZI) mesh contains 20 MZIs and 48 phase shifters. Grating coupler arrays are used to couple the light between fibers and the input/output ports. Figure S1(b) presents the transmission spectrum for the referenced grating structure, the coupling efficiency at 1550 nm is about -6.9 dB and the 3 dB bandwidth is about 40 nm from 1535 nm to 1575nm. Phase shifters convert a change in power dissipated across an electrical resistance to a change in waveguide refractive index by the thermo-optic effect. The imparted phase dependent on the applied voltage (V) may be expressed as [30]

$$\theta = \frac{2\pi V^2}{T}. \tag{S1}$$

MZIs with internal phase shifters are used to convert the phase change to the output power. The measured power distribution dependent on the square of applied voltage is presented in Fig. S1(c). The measured average period of T is about $1.07\times10^8\, mV^2$. The phase tuning efficiency is measured to be 27 mW per π phase shift and the electrical resistance is about 2000 Ω.

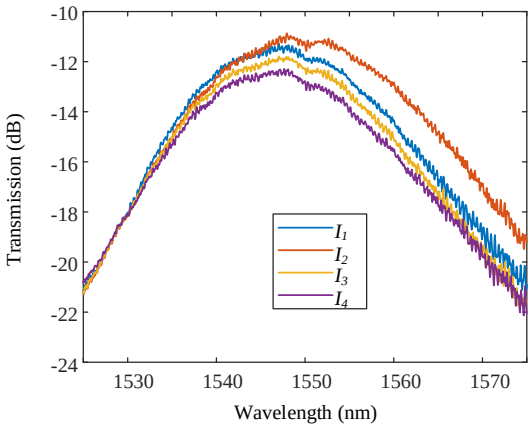

Fig. S2. Chip loss characterization. The total transmission spectra by summing the power from all the output ports in Part (3) and Part (4) are measured when the light is injected from different input ports.

The transmission spectra of the chip shown in Fig. S2, are measured by summing the output power of all the output ports in Part (3) and Part (4), when the light is injected from different input ports (labeled by I_1, I_2, I_3, I_4). The average transmission at 1550 nm is about -12 dB.

B. *Principle of LON chip*

1) *Arbitrary matrix with coherent light*

For coherent light, the complex amplitudes of light from different paths will be linearly superimposed, thus the matrix operation is applied on the complex amplitude of light. The nanophotonic processor is programmed by setting the voltages on the internal and external phase shifters of each MZI. The MZI shown in Fig. S1a implements a 2 × 2 unitary transformation on the input state of the form,

$$U_{MZIn} = R(n) = \frac{1}{2}\begin{bmatrix} e^{i\alpha_n}(e^{i\theta_n}-1) & ie^{i\alpha_n}(e^{i\theta_n}+1) \\ ie^{i\beta_n}(e^{i\theta_n}+1) & e^{i\beta_n}(1-e^{i\theta_n}) \end{bmatrix}. \tag{S2}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2943347, IEEE Journal of Selected Topics in Quantum Electronics

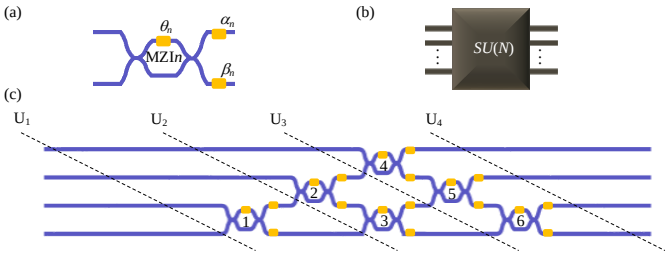

Fig. S3. Diagram of the nanophotonic processor. (a) Schematic illustration of a single phase shifter in the MZI and the transmission curve for tuning the internal phase shifter. (b) Sketch map of an *SU*(*N*) core. (c) Mesh and MZI Numbering scheme of *SU*(*4*)

All the phase are normalized to the range of $[0, 2\pi)$. A *SU*(*N*) core shown in Fig. S3(b), namely a unitary matrix with rank *N*, can be decomposed into sets of *SU*(2) rotations implemented by cascaded programmable MZIs [19]. Figure S3(c) presents an example of *SU*(*4*) mesh. It contains 6 MZIs and 18 thermo-optic phase shifters. The relationship in different planes can be written by

$$\begin{aligned} U_2 &= R_{1,1}U_1 = \begin{bmatrix} 1 & & \\ & 1 & \\ & & R(1) \end{bmatrix} U_1 \\ U_3 &= R_{2,1}R_{2,2}U_2 = \begin{bmatrix} 1 & & \\ & 1 & \\ & & R(3) \end{bmatrix} \begin{bmatrix} 1 & & \\ & R(2) & \\ & & 1 \end{bmatrix} U_2 \\ U_4 &= R_{3,1}R_{3,2}R_{3,3}U_3 \\ &= \begin{bmatrix} 1 & & \\ & 1 & \\ & & R(6) \end{bmatrix} \begin{bmatrix} 1 & & \\ & R(5) & \\ & & 1 \end{bmatrix} \begin{bmatrix} R(4) & & \\ & 1 & \\ & & 1 \end{bmatrix} U_3 \end{aligned} \quad \text{(S3)}$$

The final *SU*(*4*) is given by

$$SU(4) = R_{3,1}R_{3,2}R_{3,3}R_{2,1}R_{2,2}R_{1,1} \,. \quad \text{(S4)}$$

The same analysis can be applied on *SU*(*N*) core. A unitary matrix with rank *N* can be generally written as a product of *N*(*N* - 1)/2 rotation matrices

$$SU(N) = R_{N-1,1}R_{N-1,2}\cdots R_{N-1,N-1}\cdots R_{3,1}R_{3,2}R_{3,3}R_{2,1}R_{2,2}R_{1,1} \,. \quad \text{(S5)}$$

To implement any matrix, we first note that a general matrix (*M*) can be decomposed as $M = U\Sigma V$ through singular value decomposition (SVD), where *U* and *V* are both unitary matrices, Σ is a diagonal matrix with his eigenvalues on the diagonal. So the structure in Fig. 1a has the ability to implement any 4x4 matrix.

2) Nonnegative matrix with incoherent light

For incoherent light, the complex amplitudes of light from different paths will be still linearly superimposed at etch time, but in the case the light from different sources have random phase noise, making the collected output power is a time-averaging value. Here, we analyze the incoherent light transmission using the 4x4 matrix. In the form of complex amplitude, the transmission equation is given by

$$\begin{aligned} U_4 = MU_1 &= \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \\ &= u_1 \begin{bmatrix} m_{11} \\ m_{21} \\ m_{31} \\ m_{41} \end{bmatrix} + u_2 \begin{bmatrix} m_{12} \\ m_{22} \\ m_{32} \\ m_{42} \end{bmatrix} + u_3 \begin{bmatrix} m_{13} \\ m_{23} \\ m_{33} \\ m_{43} \end{bmatrix} + u_4 \begin{bmatrix} m_{14} \\ m_{24} \\ m_{34} \\ m_{44} \end{bmatrix} \end{aligned} \,. \quad \text{(S6)}$$

Here m_{ij} can be any complex value. The collected output powers are given by

$$I = \iint |U_4|^2 dt \,. \quad \text{(S7)}$$

Since u_1, u_2, u_3 and u_4 have random phase noise, all the cross terms ($\iint u_i u_j^* dt$, $i \neq j$) from different sources in equation (S7) will be zeroes, so the final output powers are given by

$$\begin{aligned} I &= \iint |u_1|^2 dt \begin{bmatrix} |m_{11}|^2 \\ |m_{21}|^2 \\ |m_{31}|^2 \\ |m_{41}|^2 \end{bmatrix} + \ldots + \iint |u_4|^2 dt \begin{bmatrix} |m_{14}|^2 \\ |m_{24}|^2 \\ |m_{34}|^2 \\ |m_{44}|^2 \end{bmatrix} \\ &= = \begin{bmatrix} |m_{11}|^2 & |m_{12}|^2 & |m_{13}|^2 & |m_{14}|^2 \\ |m_{21}|^2 & |m_{22}|^2 & |m_{23}|^2 & |m_{24}|^2 \\ |m_{31}|^2 & |m_{32}|^2 & |m_{33}|^2 & |m_{34}|^2 \\ |m_{41}|^2 & |m_{42}|^2 & |m_{43}|^2 & |m_{44}|^2 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \end{bmatrix} \end{aligned} \,. \quad \text{(S8)}$$

Obviously, the matrix operation is applied on the power of light for the incoherent light, and the transmission matrix can be an arbitrary nonnegative matrix.

C. Training of multiple functions with the LON chip

1) Reading and writing of input power

For each phase shifter in Part (1) of Fig. 1a, the imparted phase dependent on the applied voltage is measured with interference technique, an example is shown in Fig. S1c. The applied voltages are marked by $V_{\max}$ and $V_{\min}$ when the total power from all output ports is maximum and minimum respectively. From equation (S1), the normalized power dependent on the applied voltage is given by

$$P(\mathrm{V}) = \frac{1}{2}\left[1 - \cos\left(\frac{V^2 - V_{\min}^2}{V_{\max}^2 - V_{\min}^2}\pi\right)\right]. \quad \text{(S9)}$$

And the applied voltage dependent on normalized input power is given by

$$V = \sqrt{\frac{V_{\max}^2 - V_{\min}^2}{\pi}\arccos(1-2P) + V_{\min}^2} \,. \quad \text{(S10)}$$

Equation (S9) and equation (S10) are used to read and write the input vector respectively.

2) Numerical gradient descent algorithm

For certain targeted function, the cost function (CF) is vital for a successful training. All the training will proceed to make the CF best. A numerical gradient descent algorithm is used to

train the network. Theoretically, the training needs to combine forward propagation and backward propagation methods similar to the deep learning [30]. The forward propagation is used to calculate the output as the data for the next iteration and then the backward propagation is aimed to estimate the errors and find the gradient descent. This training algorithm is also called by gradient descent algorithm, which is a common method for training artificial neural networks (ANNs). In our design, the LON chip can output automatically and timely, provided the input is set. And the gradient descent can be alternatively measured by fine tuning each parameter. So in our design, no backward propagation is needed and the forward propagation can be finished by the LON itself at the speed of light. The main cost of time is limited by the output time of electronic control signal provided by external driver, the photodetection rate and the response time of the phase shifters in the LON chip. Furthermore, the LON chip can be regarded as a "black box", the internal structure of the chip is unimportant for training. The full training process is listed below:

1. Initialization: all the adjustable parameters of $\theta_i (i = 1, 2, ...)$ are set randomly.

2. Tuning each parameter: set θ_1 to $\theta_1 + \Delta\theta$ temporarily.

If $\mathrm{CF}(\theta_1 + \Delta\theta) \geq \mathrm{CF}(\theta_1)$, replace θ_1 with $\theta_1 + \Delta\theta$;

Else, replace θ_1 with $\theta_1 - \Delta\theta$.

3. Repeat Step 2 for all adjustable parameters one by one.

4. Repeat Steps 2 and 3 until the CF is converged or reach target value.

And in order guarantee the power efficiency, the first eigenvalue is fixed and equals 1, namely, the first MZI in diagonal matrix of Part (3) is always open in the training process.

D. PageRank algorithm

PageRank is the first and meanwhile the best-known algorithm used by Google Search to rank web pages in their search engine results. PageRank, named after Larry Page [7-9], is a way to measure the importance of website pages by solving the equation of $GP = P$, where G is the Google matrix recording the links of all webpages and P is the importance of all pages. The initial links of n webpages can be described by a nonnegative matrix of H, given by

$$h_{ij} = \frac{p_{ij}}{\sum_i p_{ij}}. \tag{S11}$$

Here, p_{ij} =1 and 0 mean that Page j links to Page i or not respectively. $\sum_i p_{ij}$ is the number of outbound links of Page j.

Multiple outbound links from one page to another page are treated as a single link and links from a page to itself are ignored. If a page has no outbound links to other pages, it becomes a dangling node and will exhaust all the importance in the iterative process, resulting in that the final PageRank scores of website pages will be all zeroes. To avoid this, pages with no outbound links are assumed to link out to all pages evenly in the collection, namely, all p_{ij} (h_{ij}) are normalized to 1 ($1/n$) when Page j is dangling. Similarly, there will be a sink appearing in links of webpages, if multiple pages have no outbound links to the other pages. The importance of pages outside the sink will be exhausted in the iterative process. To revise the PageRank scores, random transitions are added to all nodes in the Web. The probability of random transitions is given by 1-d, then the final Google matrix is given by

$$G = dH + \frac{1-d}{n} J. \tag{S12}$$

Here, all the elements of J are equal to one and d is called by damping factor. Various studies have tested different damping factors, but it is generally assumed that the damping factor will be set around 0.85[7]. The final PageRank scores of P can be obtained by

$$GP = P. \tag{S13}$$

REFERENCES

[1] K. A. Gallivan, M. Heath, E. Ng, B. Peyton, R. Plemmons, C. Romine*, et al.*, *Parallel algorithms for matrix computations* vol. 22: Siam, 1990.

[2] D. S. Watkins, *Fundamentals of matrix computations* vol. 64: John Wiley & Sons, 2004.

[3] G. H. Golub and C. F. Van Loan, *Matrix computations* vol. 3: JHU Press, 2012.

[4] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature,* vol. 401, p. 788, 10/21/online 1999.

[5] F. Sha and L. K. Saul, "Real-time pitch determination of one or more voices by nonnegative matrix factorization," in *Advances in Neural Information Processing Systems*, 2005, pp. 1233-1240.

[6] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the national academy of sciences,* vol. 101, pp. 4164-4169, 2004.

[7] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems,* vol. 30, pp. 107-117, 1998.

[8] M. Richardson and P. Domingos, "The intelligent surfer: Probabilistic combination of link and content information in pagerank," in *Advances in neural information processing systems*, 2002, pp. 1441-1448.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab1999.

[10] J. W. Goodman, A. R. Dias, and L. M. Woody, "Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms," *Opt. Lett.,* vol. 2, pp. 1-3, 1978/01/01 1978.

[11] S. F. Habiby and S. A. Collins, "Implementation of a fast digital optical matrix–vector multiplier using a holographic look-up table and residue arithmetic," *Appl. Opt.,* vol. 26, pp. 4639-4652, 1987/11/01 1987.

[12] R. A. Athale and W. C. Collins, "Optical matrix-matrix multiplier based on outer product decomposition," *Appl. Opt.,* vol. 21, pp. 2089-90, Jun 15 1982.

[13] R. P. Bocker, S. R. Clayton, and K. Bromley, "Electrooptical matrix multiplication using the twos complement arithmetic for improved accuracy: erratum," *Appl. Opt.,* vol. 22, pp. 3149-3149, 1983/10/15 1983.

[14] S. Cartwright, "New optical matrix–vector multiplier," *Appl. Opt.,* vol. 23, pp. 1683-1684, 1984/06/01 1984.

[15] J. Hong and P. Yeh, "Photorefractive parallel matrix–matrix multiplier," *Opt. Lett.,* vol. 16, pp. 1343-1345, 1991/09/01 1991.

[16] C. Gu, S. Campbell, and P. Yeh, "Matrix–matrix multiplication by using grating degeneracy in photorefractive media," *Opt. Lett.,* vol. 18, pp. 146-148, 1993/01/15 1993.

[17] B. Liu, L. Liu, L. Shao, and H. Chen, "Matrix–vector multiplication

in a photorefractive crystal," *Opt. Commun.,* vol. 146, pp. 34-38, 1998/01/15/ 1998.

[18] S. Mukhopadhay, D. N. Das, P. P. Das, and P. Ghosh, "Implementation of all-optical digital matrix multiplication scheme with nonlinear material," *Optical Engineering,* vol. 40, pp. 1998-2003, 2001.

[19] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Phys. Rev. Lett,* vol. 73, pp. 58-61, Jul 4 1994.

[20] A. Annoni, E. Guglielmi, M. Carminati, G. Ferrari, M. Sampietro, D. A. B. Miller*, et al.*, "Unscrambling light—automatically undoing strong mixing between modes," *Light: Sci. Appl.,* vol. 6, p. e17110, 2017.

[21] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, "Optimal design for universal multiport interferometers," *Optica,* vol. 3, pp. 1460-1465, Dec 20 2016.

[22] T. W. Hughes, M. Minkov, Y. Shi, and S. H. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica,* vol. 5, pp. 864-871, Jul 20 2018.

[23] P. L. Mennea, W. R. Clements, D. H. Smith, J. C. Gates, B. J. Metcalf, R. H. S. Bannerman*, et al.*, "Modular linear optical circuits," *Optica,* vol. 5, pp. 1087-1090, Sep 20 2018.

[24] D. A. B. Miller, "Perfect optics with imperfect components," *Optica,* vol. 2, p. 747, 2015.

[25] A. Ribeiro, A. Ruocco, L. Vanacker, and W. Bogaerts, "Demonstration of a 4 × 4-port universal linear circuit," *Optica,* vol. 3, pp. 1348-1357, 2016/12/20 2016.

[26] D. A. B. Miller, "Self-configuring universal linear optical component [Invited]," *Photon. Res.,* vol. 1, pp. 1-15, 2013.

[27] D. Perez, I. Gasulla, F. J. Fraile, L. Crudgington, D. J. Thomson, A. Z. Khokhar*, et al.*, "Silicon Photonics Rectangular Universal Interferometer," *Laser Photon. Rev.,* vol. 11, p. 1700219, 2017.

[28] D. Perez, I. Gasulla, L. Crudgington, D. J. Thomson, A. Z. Khokhar, K. Li*, et al.*, "Multipurpose silicon photonics signal processor core," *Nat. Commun.,* vol. 8, p. 636, Sep 21 2017.

[29] L. M. Zhuang, C. G. H. Roeloffzen, M. Hoekman, K. J. Boller, and A. J. Lowery, "Programmable photonic signal processor chip for radiofrequency applications," *Optica,* vol. 2, pp. 854-859, Oct 20 2015.

[30] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg*, et al.*, "Deep learning with coherent nanophotonic circuits," *Nat. Photon.,* vol. 11, pp. 441-446, Jul 2017.

[31] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi*, et al.*, "All-optical machine learning using diffractive deep neural networks," *Science,* vol. 361, pp. 1004-1008, Sep 7 2018.

[32] S. Ahmed, "Matrix Algebra: Theory, Computations, and Applications in Statistics," *Technometrics,* vol. 50, p. 238, 2008.

[33] H. Zhou, S. Yan, Y. Wei, Y. Zhao, Z. Cheng, J. Qie*, et al.*, "Silicon-based polarization analyzer by polarization-frequency mapping," *APL Photonics,* vol. 3, p. 106105, 2018.

[34] S. Yan, X. Zhu, L. H. Frandsen, S. Xiao, N. A. Mortensen, J. Dong*, et al.*, "Slow-light-enhanced energy efficiency for graphene microheaters on silicon photonic crystal waveguides," *Nat. Commun.,* vol. 8, p. 14411, Feb 9 2017.

[35] L. Ke, Y. C. Ran, K. Sikandar, and S. V. J., "Review and perspective on ultrafast wavelength-size electro-optic modulators," *Laser Photon. Rev.,* vol. 9, pp. 172-194, 2015.

[36] N. Daldosso, M. Melchiorri, F. Riboli, M. Girardini, G. Pucker, M. Crivellari*, et al.*, "Comparison among various Si/sub 3/N/sub 4/ waveguide geometries grown within a CMOS fabrication pilot line," *J. Lightwave Technol.,* vol. 22, pp. 1734-1740, 2004.

[37] W. Liu, M. Li, R. S. Guzzon, E. J. Norberg, J. S. Parker, M. Lu*, et al.*, "A fully reconfigurable photonic integrated signal processor," *Nat. Photon.,* vol. 10, pp. 190-195, 2016.

[38] L. Vivien, A. Polzer, D. Marris-Morini, J. Osmond, J. M. Hartmann, P. Crozat*, et al.*, "Zero-bias 40Gbit/s germanium waveguide photodetector on silicon," *Opt. Express,* vol. 20, pp. 1096-1101, 2012/01/16 2012.

Hailong Zhou received the Ph.D. degree in optoelectronics engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2017. He is currently a postdoctor with Wuhan National Laboratory for Optoelectronics, HUST. He is working on silicon photonics, inverse design and so forth.

Yuhe Zhao is currently working toward the Ph.D. degree at the Huazhong University of Science and Technology, Wuhan, China. His current research interests include photonic generation of microwave signals, photonic microwave frequency identification, and silicon photonics.

Gaoxiang, Xu is currently working toward the Ph.D. degree at the Huazhong University of Science and Technology, Wuhan, China. His current research interests include data access in parallel file system, network storage technologies, big data storage and management.

Xu Wang is currently working toward the Ph.D. degree at the Huazhong University of Science and Technology, Wuhan, China. His current research interests include photonic generation of microwave signals, photonic microwave frequency identification, and silicon photonics.

Zhipeng Tan received the Ph.D. degree in school of computing from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2008. He is currently a Full Professor with Wuhan National Laboratory for Optoelectronics, HUST. He is working on data access in parallel file system, network storage technologies, big data storage and management.

Jianji Dong received the Ph.D. degree in optoelectronics engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2008. From November 2008 to February 2010, he was with the Centre of Advanced Photonics and Electronics, Cambridge University, Cambridge, U.K., as a Research Associate. He is currently a Full Professor with Wuhan National Laboratory for Optoelectronics, HUST. He is working on silicon photonics and integrated microwave photonics. He has made some contributions in broad areas, such as ultrafast photonic differentiator using silicon waveguides, microwave signal processing, and so forth.

Xinliang Zhang received the Ph.D. degree in physical electronics from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2001. He is currently with Wuhan National Laboratory for Optoelectronics and the School of Optical and Electronic Information, HUST, as a Professor. His current research interests include InP-based and Si-based devices and integration for optical network, high-performance computing, and microwave photonics. In 2016, he was elected as an OSA Fellow.