IEEE Photonics Journal
An IEEE Photonics Society Publication
Open Access

Revision: PJ-007964-2018

Optical versus electronic implementation of probabilistic graphical inference and experimental device demonstration using nonlinear photonics

Masoud Babaeian[1,2,*], Patrick Keiffer[1], Mark A. Neifeld[1,3], Ratchaneekorn Thamvichai[3], Robert A. Norwood[1], Pierre-A. Blanche[1], John Wissinger[1], and N. Peyghambarian[1]

[1]College of Optical Sciences, University of Arizona, Tucson, AZ 85721, USA
[2]Department of Physics, University of Arizona, Tucson, AZ 85721, USA
[3]Electrical and Computer Engineering, University of Arizona, Tucson, AZ 85721, USA
*Babaeian@physics.arizona.edu

Abstract: The probabilistic inference model has been widely used in various areas such as error-control coding, machine learning, speech recognition, artificial intelligence and statistics. In this paper, we study both computation and communications power consumption of optical-based and electronic-based implementations of the probabilistic inference algorithm used in solving large scale problems. Our analysis indicates that the optical implementation provides substantial reduction for power and area compare to the electronic-based solutions as problems become large. For a network with 1 million nodes and 100 alphabet size, our proposed wavelength multiplexed all-optical implementation requires approximately 200 kilowatts (kW) of power as compared with 1.47 gigawatts (GW) and 1.7 megawatts (MW) using CPU-based and sub-threshold VLSI-based systems, respectively. The optical-based solution is tolerant to shot noise and imperfections of optical modules used in the architecture as well. We also performed an all-optical experimental verification of a graphical inference as the proof of concept and have demonstrated the essential mathematical operations, multiplication and normalization (division), in photonics operations using nonlinear bulk materials. The normalization and multiplication are shown optically through a pump-probe saturation process and a logarithm-summation-exponential (log-sum-exp) operation, respectively. We used single mode silicon waveguide (SiW) and single-wall carbon nanotube (SWCNT) as nonlinear optical materials to implement logarithm and exponential operations respectively. The SWCNT is also used as the nonlinear component in the pump-probe saturation experiment to implement the normalization function.

1. Introduction

1.1 Background and Motivation

Today's computer architecture has reached a consensus: electronics is superb at processing data and optics is excels at transporting the information. This statement finds its embodiment in the modern supercomputer and data center layout where semiconductor transistors handle the digital bits and optical fibers interconnect the processors carrying those same bits as photons. The cost of communications is the rate of energy required for electronic-optical-electronic (EOE) and optical-electronic-optical (OEO) conversion. For large scale problems, when large amount of data need to be transferred between central processing units (CPUs), and for larger distances, the power consumption is dominated by the communications cost instead of the computation [1-3]. Optical computing in the last 30 years has been able to demonstrate Fourier transform [4] and some other mathematical operations such as vector

Digital Object Identifier 10.1109/JPHOT.2018.2871822

matrix multiplication [5,6] and matrix inversion [7-9]. However, these types of optical computations have shown limited applications due to their lack of versatility and scalability: optical processors are difficult to reconfigure in order to solve different problems. Additionally, the size of optical components, even in integrated optics, is generally much larger than electronic transistors. The versatility issue can be tackled by introducing a computational method that requires no change in the algorithm and only changes in the parametric components for solving different problems. In this case, a fully optical solution can be implemented and applied to a large variety of problems by simply reconfiguring the input. One such technique is probabilistic graphical inference (PGI) [10]. PGI is an extremely powerful method for computing joint probability distributions over a large number of random variables that interact with each other [10-14]. This technique has found practical applications in a wide variety of fields such as artificial intelligence [15,16], machine learning [16-19], image analysis [20] and signal processing [21]. All these applications generally work with large amount of input ranging from the million to the billions of variables. With such a large number of variables, the decomposition by the chain rule factorization is predominantly advantageous in terms of the number of operations saved. The popularity of PGI is largely due to the emergence of big data in diverse disciplines, from medicine to economics to social networks [22-24].

1.2 Graphical Model and the Message Passing Algorithm

The graphical model used to represent a PGI problem is composed of two fundamental elements: the structure and the parametric components. The structure is the layout of the graph where nodes corresponding to the random variables are connected by edges representing the conditional dependencies. The parametric components encode the state of the node x into a probability function $P(x)$ over an alphabet K, as well as the conditional probability distribution of the edge linking y to x: $P(x/y)$. The alphabet K is determined by applications i.e., if the graphical model is used to represent an Ising Hamiltonian model where each node can have a spin of either -1 or +1, K is therefore 2.

There exist different types of problems that can be addressed using graphical models and the goal is to recover some information on the hidden variables based on noisy or incomplete observations. In the case of calculating the marginal distribution for unobserved variables, belief propagation, also known as the sum-product message passing algorithm (SPMPA), is particularly effective [12]. This iterative algorithm works by passing a "message" ($\mu_{i\to a}$) that contains the "influence" that node i is exerting on node a. This message is computed by the product of the probability vector of node i with the conditional dependency between a and i. When node a is connected to several nodes, the message sent to a is the product of the messages from all neighbor nodes of a. In other words, the estimated marginal distribution of each individual node is proportional to the product of all messages from its neighbor nodes

$$p(x_a) = \frac{1}{Z}\prod_i \mu_{i\to a}(x_a), \tag{1}$$

where Z is a normalization factor to ensure that the result is a probability vector: $p(x_a) \in [0,1]$ and $\sum_{x_a} p(x_a) = 1$ (see Supporting Information section 1). After several iterations in which the state of the nodes is updated accordingly, the value at each node eventually converges towards the marginal distribution (exactly if the graph is acyclic).

The SPMPA for a node, i^{th} node, can be illustrated in Figure 1a where node i is assumed to connect to j neighboring nodes. For a fully connected graph with N nodes, each node has N-1 neighbor nodes. Each neighbor node sends its message by first multiplying its probability vector with the compatibility matrix

$$Y_{ij} = C_{ij}X_j(k), \tag{2}$$

where *i* is the receiving node and *j* is the node sending the message, $j \neq i$. $X_j(k)$ is the probability vector of node *j* at time *k*. C_{ij} is a compatibility matrix between node *i* and *j*. This operation is called vector-matrix-multiplication (VMM). The messages from all neighbor nodes are then multiplied together as

$$Z_i(k) = \prod_{m \neq i, m=1}^{j} Y_m(k) \tag{3}$$

Eq. (3) is an element-wise product of all incoming messages. The product is then normalized

$$X_i(k+1) = \mathrm{Norm}\big(Z_i(k)\big). \tag{4}$$

To ensure that the sum of all its elements equals to 1, representing the updated probability vector of node *i*. These operations are applied to every node in order to update its probability vector. The updated vectors are then used in the following iterations until their values reach the steady state. The final and stable probability vector is then used to decide the potential state of the node. It would be easy to do the multiplication optically in Eq. (3) using a combination of logarithmic-summation-exponential (log-sum-exp) operations [25,26]

$$\prod_{m=1}^{j} Y_m = \exp(\sum_{m=1}^{j} \mathrm{Log}(Y_m)) \tag{5}$$

Figure 1b shows the algorithm for node i^{th} implemented using the log-sum-exp operations instead of direct multiplication. Implementing the operations presented in Figure 1b with optics induces noise. To investigate the effect of this noise as well as graph connection density on the algorithm performance and robustness, we performed a simulation on 100-node graphs with alphabet size of *K*=2 and *K*=100. The simulation result indicates that graphs with less than 20% connection density have higher failure rate than graphs with higher connection density (see Supporting Information section 2).

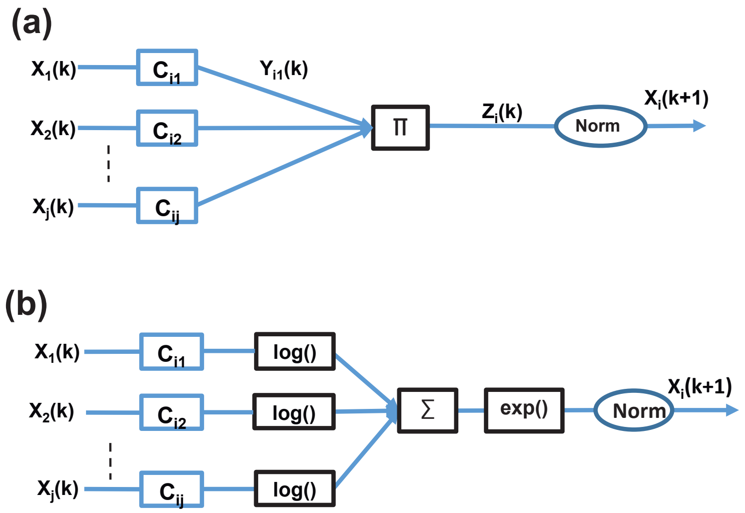

Figure 1. (**a**) SPMPA scheme for node i^{th} which is connect to j neighboring nodes. (**b**) Substitution of multiplication block in Figure 1a with a Log-sum-exp composite function according to Eq. (5).

We have performed also an electronic benchmarking of SPMPA and our analysis shows that the power consumption of computing and communication is very high for a massively parallel electronic computing system to solve a densely connected graph (e.g. each node is connected to all other nodes). As an example, for a fully connected graph (D=1) with one-million nodes, and an alphabet size of 100 (N=10^6 and K=100), the number of floating point operations for the sum-product message passing algorithm reaches roughly ~ 10^{18}. To solve this problem for a 20% graph connectivity (D=0.2) using electronic computing platforms on the CPU and sub-threshold very large scale integration (VLSI), we find that the total power requirement for computation and communication is ~1.47 GW and 1.7 MW respectively (see Supporting Information section 3). These powers budgets are dominated by communication power consumption.

2. Optical Solution

2.1 Wavelength Multiplexing

Possible implementations of an optical solutions for the SPMPA has revealed that a wavelength multiplexing approach is the most favorable solution in term of minimizing both power and the number of components [25-27]. This is because the number of nodes is by far the largest quantity in the graphical architecture. In this multiplexing layout, the spectral bandwidth is equally divided and used for each node as presented in Figure 2a. The SPMPA scheme in Figure 2a contains N nodes and the alphabet size is K. Each node has a probability vector of size K where each element is a probability that the node has alphabet 0, 1, …, K -1, respectively. If there is no prior information on a node, each element of its probability vector is equal to 1/K. To determine the updated probability vector of a specific node like node m in Figure 2a, each probability vector from its neighbor nodes is multiplied by a matrix whose elements are conditional probabilities. The output vectors (messages from all neighboring nodes) are then multiplied element-wise and normalized to yield the updated probability vector of that specific node. The multiplication of all messages is substituted with a composite function of log-sum-exp as discussed in Eq. (5).

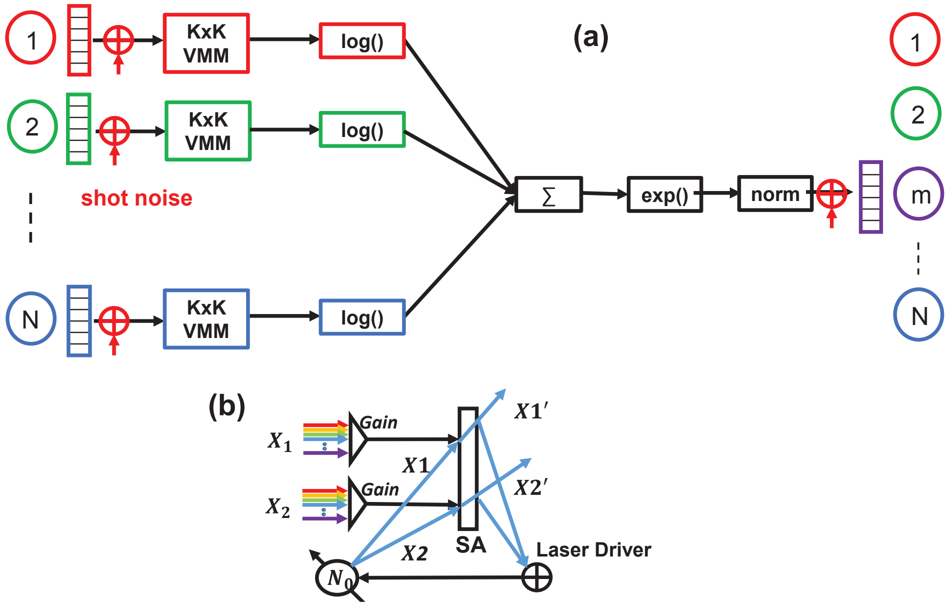

Figure 2. (**a**) Graphical scheme of the SPMPA for node m. Effect of optical shot noise on SPMPA also studied to investigate the tolerance of all optical solution of graphical inference problem. Each node has different wavelength as it is indicated with different color. (**b**) Normalization set-up to normalize two numbers and concept of wavelength remapping. Each element of the probability vector is modulated in the presence of a broadband pump which requires spatial separation in the saturable absorber. An electronic feed-back-loop system adjusts the power of the output probe source such that $X1' + X2'$ remains constant.

These mathematical operations are enforced to every node in order to get its updated probability vector. The updated probability vectors are then used in the consecutive iterations until their values converge. The final and stable probability vector is used to decide the potential alphabet that a node has.

2.2 VMM

Persistent spectral hole burning (PSHB) can be used to implement the VMM [28]. Each element/pixel in the K×K PSHB plane can be altered such that its absorption is changed according to element values of the compatibility matrix C_{ij}. The output from the PSHB medium therefore corresponds to multiplying C_{ij} with the probability vector where each element of the vector is represented by a specified light intensity. The mathematical model used for PSHB is

$$I_{out} = I_0 \exp(-\alpha_0(1-\eta)L), \tag{6}$$

where I_0 is the input irradiance, α_0 is linear absorption coefficient, η is hole depth and L is the thickness of material. For the parameter α_0, 400/m (Eu-YSO materials) [29] is used and η is chosen to be 0 or 0.5. Our initial study chose elements of an ideal VMM, a $K \times K$ matrix, to be 0.9 and 0.1 (e.g. [0.9 0.1; 0.1 0.9]) for K = 2. This is to ensure that the neighbor node will likely send the correct message to the target node in a noisy environment. It is obvious from Eq. (6) that we cannot achieve the two values of 0.9 and 0.1 with the chosen parameters α_0 and η. We then conducted a study such that by choosing L = 5.7 mm, our model of the SPMPA still functions. Thus, the term $exp(-\alpha_0(1-\eta)L)$ gives the values 0.3198 and 0.1023 for η = 0.5 and 0, respectively.

2.3 Logarithm

Two-photon-absorption (TPA) can be used to approximately represent the logarithm operation [30]. This can be implemented in silicon waveguides. The following equation is used in the simulation model

$$I_{out} = \frac{I_0 \exp(-\alpha_0 L)}{1 + C_{TPA} I_0}, \tag{7}$$

where I_{out} is the output irradiance, $C_{TPA} = \beta\,[1 - exp(-\alpha_0 L)]/\alpha_0$ and β is the two photon absorption coefficient [31]. The photon number instead of the light irradiance is used in the simulation in order to study the effect of shot noise. The TPA parameter, C_{TPA}, is chosen to be $2/N_0$ where N_0 is initial number of photons for each node.

2.4 Exponential

A saturable absorber (SA) can be used to approximately represent the exponential operation [26]. This can be implemented using carbon nanotubes as SAs. The following equation is used in the simulation model

$$I_{out} \exp\left({}^{I_{out}}/_{I_{sat}}\right) = I_0 \exp\left({}^{I_0}/_{I_{sat}} - \alpha_0 L\right), \tag{8}$$

where I_{sat} is saturation irradiance [32]. Photon number instead of the irradiance is used in the simulation and the parameters of SA for the exponential module, $\alpha_o L$ and N_{sat},are chosen to be 1. Note that the parameters of the TPA and SA modules are studied and chosen such that the optical-based algorithm yields no failure rate when no noise is added.

2.5 Normalization

A SA can be used for the normalization. In our case, the normalization module is responsible for two functions: a) make the sum of all elements of each normalized probability vector remain constant, and b) integrate over the input spectrum and translate to a node-specific output wavelength [26,27]. Figure 2b shows a simple optical model for normalization of two values. N_0 is adjustable so that $X1' + X2'$ remains constant. The following equations are used in the model

$$X1' = N_0 \exp\left(\frac{N_0 + X_1.Gain}{N_s} - \alpha_0 L\right), \tag{9a}$$

$$X2' = N_0 \exp(\frac{N_0 + X_2.Gain}{N_s} - \alpha_0 L), \tag{9b}$$

where N_s is the saturation photon number value and *Gain* is used so that the amplified input values can saturate the materials. The parameters used in the simulation are $\alpha_0 L$ = 1, $I_o = N_o, N_s = 10N_o$, and *Gain* = 224 for an alphabet size *K* = 100. The values are studied and selected to ensure that the message passing model with these normalization parameters still performs correctly. We study two cases: 1) an ideal normalization where $X1'$ and $X2'$ are equal to $N_0(X_1/(X_1 + X_2))$ and $N_0(X_2/(X_1 + X_2))$, respectively; and 2) a normalization device using SA as discussed in Eq. (9a) and Eq. (9b). For logarithm, exponential, VMM and normalization operations the photon number is used in the simulation in order to study the effect of shot noise. The optical based solution of SPMPA yields 1% failure rate in presence of shot noise which indicates the optical solution of SPMPA is very tolerant and robust. Our theoretical analysis of power consumption of SPMPA in optical domain using wavelength multiplexing architecture shows 200 kW (N=10^6, K=100 and D=20%) which in principle is a great advantage over the electronic computing platforms (see Supporting Information section 4 and 5).

3. Experimental device demonstration

To demonstrate the possibility of an all optical implementation of the SPMPA, we have performed the essential mathematical operations: multiplication and normalization, in nonlinear optical bulk materials. We used single mode silicon waveguides (SiWs) for TPA, and single-wall carbon nanotubes (SWCNTs) as the SA. For each experiment we start with a simulation that describes each element's mathematical behavior.

3.1 Multiplication

A numerical solution of Eq. (8) and its fit with an exponential function are plotted in terms of P_{out} versus P_{in} in Figure 3a, where P_{in} and P_{out} are input and output average power, respectively. Note that we can use average power or photon number instead of peak irradiance without any change in the mathematical analog concepts. It has to be noted that we have to limit the range of the fit in order to get a maximum overlap between the numerical solution of the SA equation and the desired exponential function. This reduces the dynamic range of the mathematical operation. Figure 3b shows the numerical solution of Eq. (7) and its fit with a logarithm function in terms of P_{out}versus P_{in}. Likewise, bounding the dynamic range of fitting yields the maximum overlap between the numerical solution of the TPA equation and the logarithm function. Bounding the fitting range comes from the natural behavior of the TPA and SA where Eq. (7) and Eq. (8) start from zero, for no input power, whereas $\log(0)$ is undefined and $\exp(0) = 1$. Figure 3c shows the result of the combination of 5 identical logarithm inputs and an exponentiation which gives the multiplication of the inputs as it is described in Eq. (5). Ideal multiplication is plotted as a solid line in Figure 3c. The acceptable normalized-root-mean-square error (NRMSE) of fitting, which is defined as the following, should be below 1% as our simulation presented in section 1 indicated

$$\mathrm{NRMSE} = \frac{\sqrt{\langle (P_{out} - P_{fit})^2 \rangle}}{P_{max} - P_{min}} \tag{10}$$

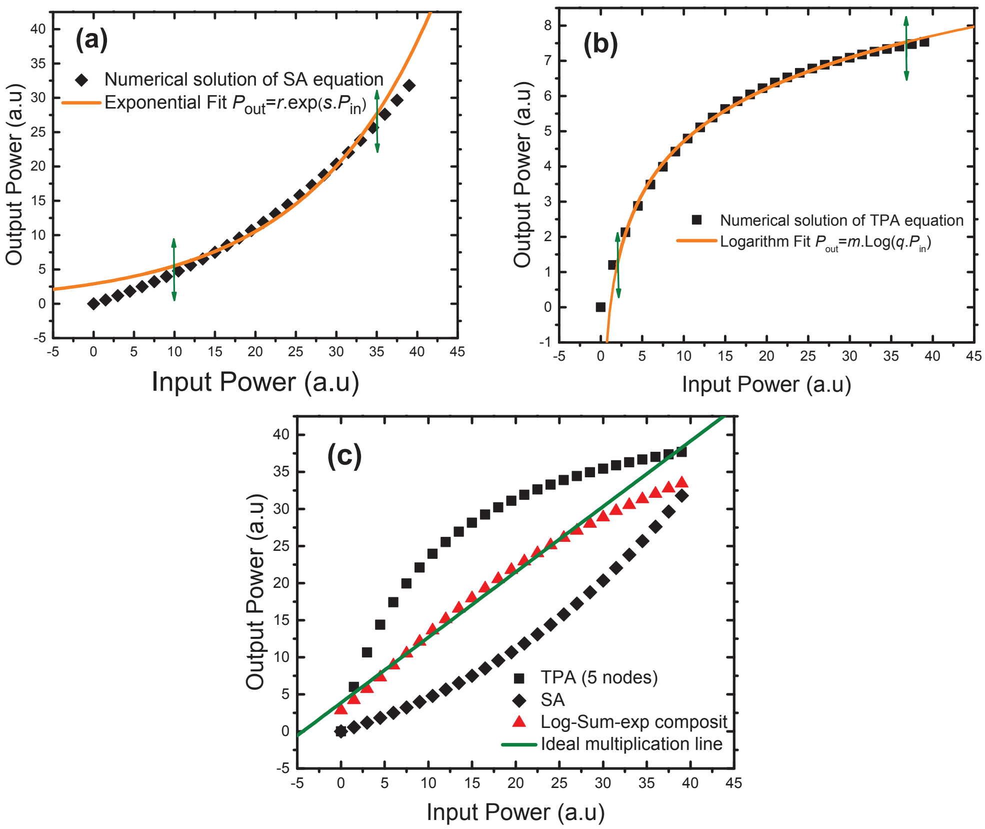

Figure 3. Numerical simulation of multiplication based on log-sum-exp combination. (a) Comparison of the SA solution (Eq. 8) with an exponential function $P_{out} = r.\exp(s.P_{in})$ where the fit coefficients are r=2.9 and s=0.064. The parameter values of the numerical simulation are $\alpha_0 = 10\ (\mathrm{a.u}), P_{sat} = 49\ (\mathrm{a.u})$ and $L = 0.1\ (\mathrm{a.u})$. (b) Comparison of TPA solution (Eq. 7) with a logarithm function $P_{out} = m.\log(q.P_{in})$ where the fit coefficients are m= 2.16 and q=0.88. The parameter values of the numerical simulation are $\alpha_0 = 0.9\ (\mathrm{a.u}), L = 0.1\ (\mathrm{a.u})$ and $C_{TPA} = 0.095\ (\mathrm{a.u})$. (c) The red triangles show the composite mathematical operations of log-sum-exp for 5 inputs and the solid green line represents ideal multiplication.

We have demonstrated the multiplication experiment to multiply two average power numbers. We have used two silicon waveguides (SiWs) as TPA units and a SWCNT based fiber tapper as SA to enable logarithm and exponential functions, respectively, in the configuration presented in Figure 4a. The optical laser source that has been used for this experiment was a 1550 nm mode-locked fiber laser, producing 160 fs pulse width (at FWHM) and a 75 MHz repetition rate. As we discussed in the TPA and SA simulations, limiting the fitting range is required in order to get the logarithm and exponential functions. We used two erbium-doped fiber amplifiers (EDFAs) to amplify the output powers of the SiWs to compensate for the insertion loss from fiber-waveguide-fiber coupling. Although we utilized ultra-high numerical aperture (UHNA7) fiber for coupling the light into the SiWs the use of EDFAs were still necessary. The mode area of these waveguides (350 nm × 350 nm and 4 mm length) is also small compare to the mode area of the UHNA fibers which results in high insertion loss in the coupling ports. After each EDFA, a linear polarizer (PL) and polarizing controller (PLC) are placed in the path to insure that the output polarization result is perpendicular to the other arm's polarization. A polarizing beam combiner (PBC), which preserves the input polarization's orientation, combines the two beams with a perpendicular polarization orientation. Hence, these two beams do not interfere at the SA even though they have the same wavelength. Furthermore, a delay stage was placed in one of the arm for pulse time matching, followed by an auto-correlator at the SA with femtosecond resolution. Two variable optical attenuators (VOAs) and two beam splitters (BS1 and BS2) were used to monitor the input powers to the TPA units. Figure. 4b, Figure 4c and Figure 4d show the experimental data for P_{out} versus P_{in} and the nonlinear fit functions with the logarithm and exponential functions for the TPA units and SA block respectively. Figure 4e shows the measured output power as a result of appropriate manipulations of the two input powers, P_1 and P_2, versus the ideal multiplication of the two numbers (solid blue line). The output power has been included with two optical constants δ and ξ (Figure 4e) to take the component imperfections into account. These imperfections arise from various sources such as insertion loss of the optical components, linear and second order absorption in the SiWs and the SWCNT saturable absorber. The constants δ and ξ are related to the fit coefficients $(m_1, m_2, q_1, q_2, r, s)$ which are known and constant.

As we described in Eq. (5), the composite function of the sum of two logarithms and subsequent exponentiation yields the product of the input values. Now taking the fit coefficients from Figure 4b and Figure 4c into account, we get the summation of the two output powers from the TPA units (the PBC does the summation operation) as

$$\log(q_1.P_1) + \ m_2\log(q_2.P_2) = \log[(q_1.P_1)^{m_1} \times (q_2.P_2)^{m_2}]. \tag{11}$$

We also adjusted the gains of the two EDFAs such that m_1 became m_2 ($m_1 \approx m_2 = m$). The right-hand-side of Eq. (11) is the output of the PBC, which is the input to the SA. The SA acts on the input values based on the fit equation in Figure 4d.

$$r.\exp[s.\log((q_1.q_2P_1 \times P_2)^m)] \rightarrow r\ (q_1.q_2)^{m.s}(P_1 \times P_2)^{m.s}. \tag{12}$$

The Eq. (12) can be written as $\delta(P_1 \times P_2)^{\xi}$ where $\xi = m.s$ and $\delta = r(q_1.q_2)^{m.s}$. These coefficients ($\delta$ and ξ) embody fundamental material characteristics and all of the imperfections of the experimental set-up. Adding two gain blocks in the experimental set-up can eliminate δ and ξ and get pure mathematical multiplication of two numbers ($P_1 \times P_2$).

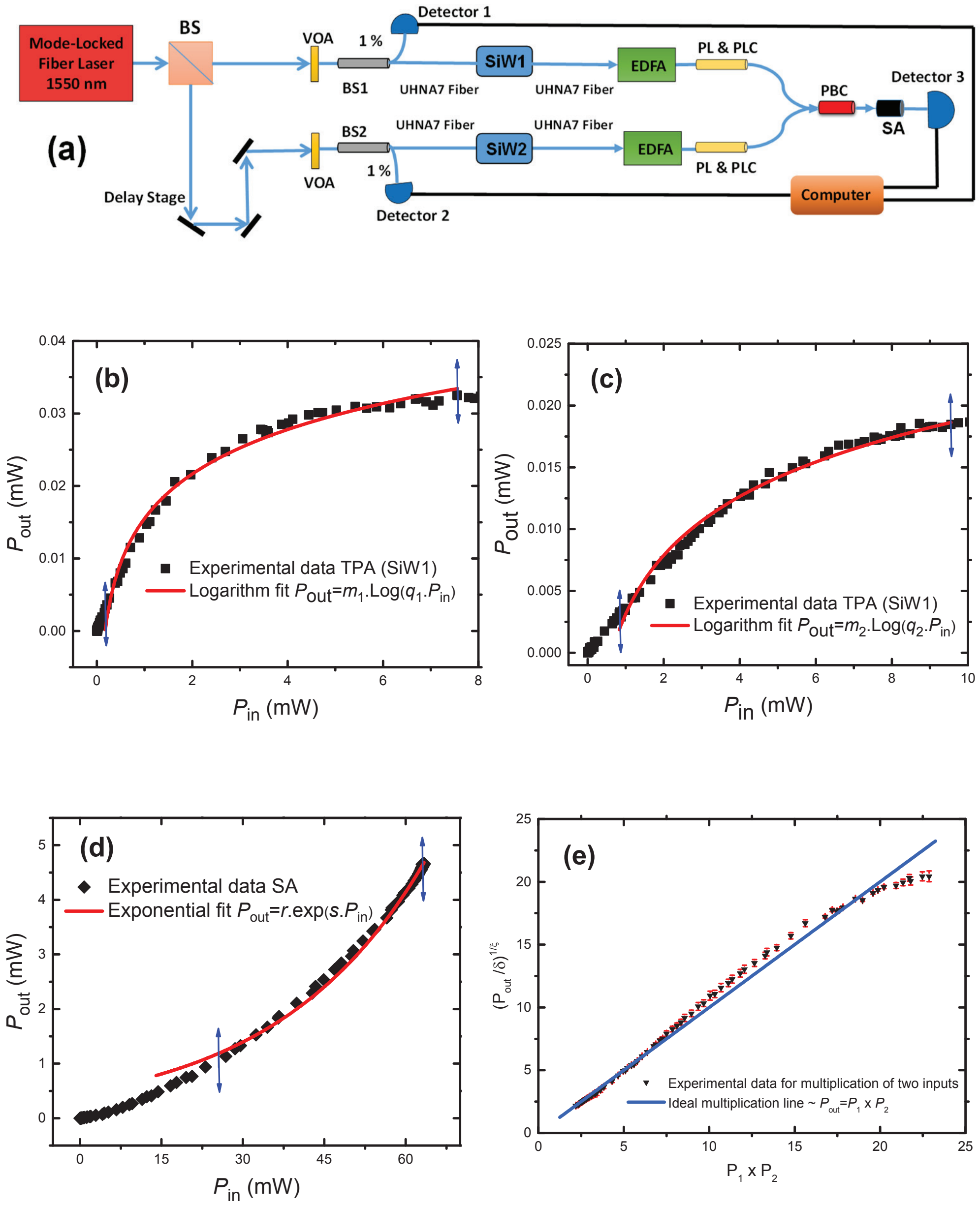

Figure 4. Experimental multiplication results. (**a**) Experimental set-up to multiply to input values. (**b**), (**c**) TPA data and the nonlinear fits $P_{out} = m_1.\log(q_1.P_1)$ and $P_{out} = m_2.\log(q_2.P_2)$ respectively. The fit coefficients are $m_1 = 0.0088$ mW, $q_1 = 5.76$ 1/mW, $m_2 = 0.0068$ mW and $q_2 = 1.57$ 1/mW. (**d**) SA data and the nonlinear fit $P_{out} = r.\exp(s.P_{in})$ where r=0.47 mW and s=0.036 1/mW. (**e**) The measured final output power (triangle) versus the ideal multiplication of the input values (solid blue line). The error bars indicate the relative percent error between ideal multiplication of two power inputs and the experimental readout.

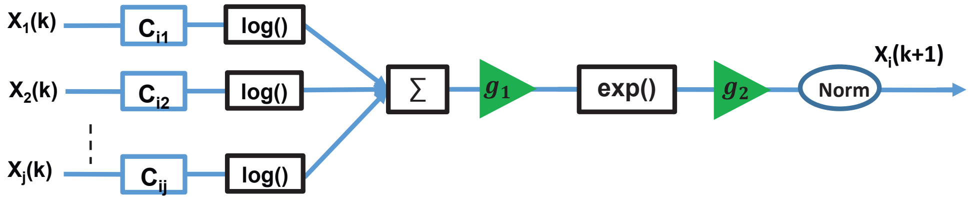

Figure 5. The modification of the multiplication unit in Figure 1b where two gain blocks are added before and after SA. This approach can conduct the value of δ and ξ to 1.

The value of the two gain units, g_1 and g_2 should be equal to $1/\xi$ and $1/(rq_1q_2)$ respectively (Figure 5). Note that if we want to multiply more than two numbers we may need to use attenuators instead of gain blocks if δ or ξ or both become greater than one. The size of the graph, density of connectivity of the graph and material characteristics dictate whether to select gain block(s) or attenuator unit(s).

3.2 Normalization and wavelength remapping

According to the wavelength multiplexing approach for the message passing algorithm implementation, the information that reaches the receiver node should be monochromatic in order to be recirculated for the next iteration. To implement normalization and wavelength translation, we propose to employ an optical pump-probe saturation experiment followed by an electrical feedback-loop system [26]. We use a SA in which by approaching the saturation intensity, we can decrease or increase the optical intensity of a probe beam. The SA also integrates over the input spectrum, and remaps the output to the proper node-specific wavelength. The pump is a broadband laser source that includes all the elements of the probability vector and the probe can be a continuous wave (CW) laser source at the node's wavelength (Figure 2b and Figure 6a). Furthermore, each element of the probability vector must be spatially separated in the SA during the normalization process. The feed-back loop controls power P_0 (average power is used instead of photon number in the experiment) in which for any value of X_1 and X_2 , $X1' + X2'$ remains constant, where $X1' = P_0\, X_1/(X_1 + X_2)$ and $X2' = P_0\, X_2/(X_1 + X_2)$ as discussed earlier. It is obvious that if the intensity of X_1 increases, the intensity of $X1'$should also increase, whereas the intensity of $X2'$ decreases suchthat $X1' + X2'$ remains constant. This effect rises from natural behavior of SA where an increase of the pump intensity causes more electrons to populate an upper energy level and, therefore, the probe light can pass the SA with less absorption.

We employed the normalization experiment as well as a simulation of an ideal normalization of two input powers X_1 and X_2 and the result of $X1' + X2' = \text{constant}$ (a.u). Figure 6c shows the experimental result for the normalization and good agreement with the simulation result provided in Figure 6b. We kept the power of laser X_2 at a constant value in experiment. as well as in the simulation. Figure 6a denotes the experimental set-up for the normalization of two inputs. The pump sources were two mode-locked femtosecond fiber lasers, and a CW laser was used as the probe. Beam splitters BS1 and BS2 monitored the input powers of X_1 and X_2 to the SAs, respectively. PBC1 and PBC2 combine the power X_1 with *X1* and power X_2 with *X2* and make them collinear at the SAs so the powers of *X1* and *X2* from the probe laser can be modulated in the presence of the pump lasers X_1 and X_2, respectively. A half-wavelength plate and a polarizer were used in one of the probe laser's path to avoid interference at detector number 3. PBC3 combines all powers, with preservation of the polarization orientation of laser *X1* and *X2* to be perpendicular, and a wavelength-division-multiplexing (WDM) splits the powers based on the pump and probe wavelength.

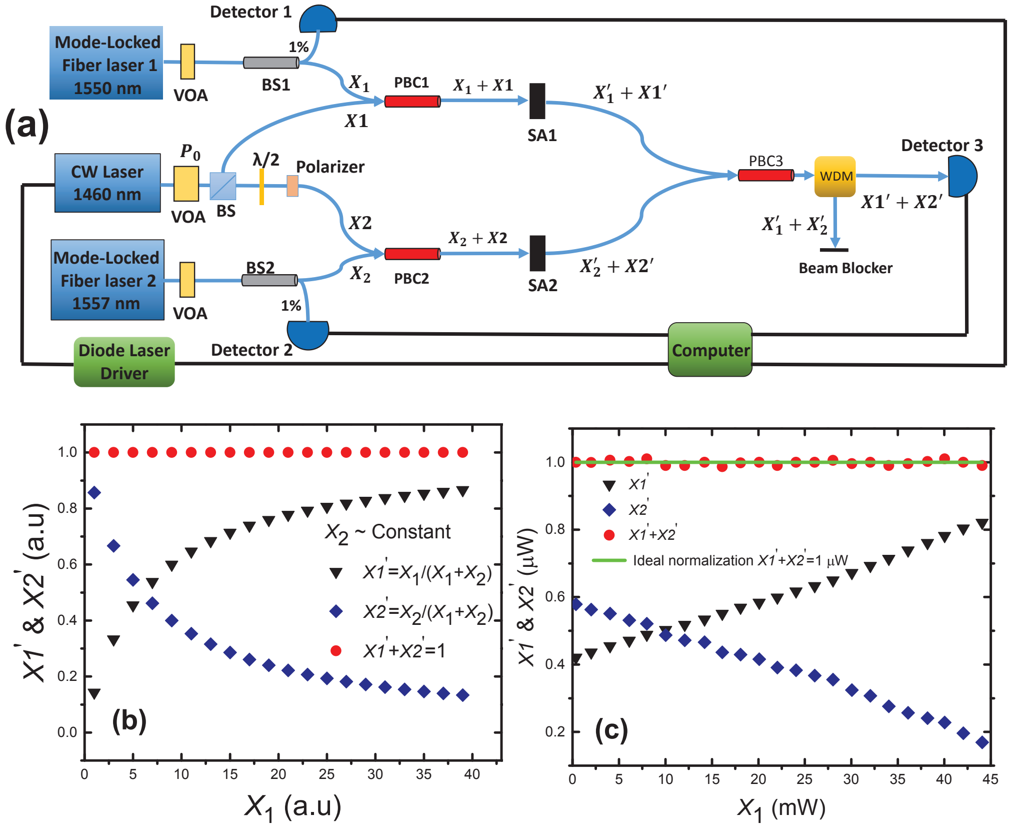

Figure 6. (a) Experimental set-up to normalize to input values as well as the wavelength remapping concept using a pump-probe saturation experiment. The probe was a CW diode laser λ_{Probe}=1460 nm. The pump sources are two mode-locked fiber lasers. The characteristics of these lasers are as follows: λ_{X_1}=1550 nm with 75 MHz repetition rate and 150 fs pulse width, and the other one λ_{X_2}=1557 nm, 8 MHz and 240 fs pulse width. (b) Simulated result to normalize two numbers X_1 and X_2 (based on ideal normalization equations) where we assume X_2 is constant. (c) Experimental result to normalize two powers where the feedback-loop system adjusts the modulated power of $X1' + X2'$ to remain constant.

4. Discussion

Considering the importance of probabilistic graphical inference (PGI) in a large number of applications, we investigated different methods for solving large-scale networks with tractable power and space requirements. Using a benchmark model composed of 1 million nodes and an alphabet of 100, we estimated that any electronic solutions would require extensive area (at least 50 m^2 of silicon), and that the power consumption would be dominated by the communications between clusters of nodes, requiring at least 1.7 MW of power using sub-threshold VLSI. This later fact put us on the path to all-optical solutions where communication power should be dramatically reduced since the photons are used for both computation and information transfer. We investigated a PGI implemented via the sum-product message passing algorithm (SPMPA), using a wavelength-multiplexing approach. Computer simulations demonstrated that this approach would be tolerant to device imperfections and noise if the graph connectivity is larger than 20%. Considering our benchmark model with a million nodes and an alphabet size of 100, the computational power and area required for the optical components was computed to be 200 kW and 57 cm^3, respectively (Table. S4, Supporting Information section 2). Therefore, optics clearly shows attractive advantages in terms of power and area compared to the available electronic platforms. Furthermore, our theoretical analysis indicates that the optical-based SPMPA's failure rate would be less than 1% in the presence of noise for each optical component, which offers a very robust solution. We should note that theoretically bandwidth of the optical coherent sources is a fundamental limitation of scaling the PGI to large number of nodes. Hypothetically, increasing the bandwidth of the coherent source, results getting higher nodes in the PGI. In the wavelength multiplexing approach, the spectral bandwidth of the coherent source needs to divide equally to denote each node in the SPMPA. Therefore, not only energy of each node reduces also the pulse width of the source expands which results lower peak power. This reduction for large number of nodes may not be enough to access TPA and SA behavior of most known materials in nature. A possible way to divide the spectral bandwidth to a large number (not millions yet) could be arrayed-waveguide grating (AWG) technology [33-35]. Also, a broadband frequency comb source (for instance, frequency combs with 800nm band, each comb having 100MHz or 0.8pm linewidth) could be a solution to produce a large number of nodes [36-41].

5. Conclusion

To establish the viability of the optical implementation, we experimentally demonstrated the two main mathematical functions required for SPMPA: multiplication (via log-sum-exp), and normalization. A single-wall carbon nanotube coated fiber taper (SWCNT) was used as a saturable absorber to implement the exponentiation, and silicon waveguides (SiWs) were used for the logarithm by means of two photon abortion. The SWCNT was also used for the normalization and wavelength remapping through a pump-probe-saturation experiment. These experiments showed that even though the dynamic range of the functions are limited due to the divergence of the optical signal near zero and at high power, the optimum range of operation is large enough to support the optical implementation of SPMPA. In this article, the nonlinear optical materials that we used (SiW and SWCNT) could be considered as "bulky". Therefore, the volume required for a million node instantiation would be much larger than the theorized 57 cm^3. However, thin film nonlinear optical materials such as graphene and $MoSe_2$ offer a possible solution to implement a large scale system-level demonstration. This embodiment is the subject of ongoing work.

Funding

Office of Naval Research (ONR) MURI program on Optical Computing under grant no. N00014-14-1-0505, National Science Foundation (NSF) ERC CIAN under grant no. EEC-0812072 and the Arizona TRIF program.

Acknowledgement

The authors would like to thank Dr. Houman Rastegarfar, and Dr. Soha Namnabat for their helpful discussions and assistance.

References

1. W. Van Heddeghem, S. Lambert, B. Lannoo, D. Colle, M. Pickavet, and P. Demeester, “Trends in worldwide ICT electricity consumption from 2007 to 2012,” Comput. Commun. **50**, 64–76 (2014).
2. D. A. B. Miller, “Device requirements for optical interconnects to silicon chips,” Proc. IEEE. **97** (7), 1166–1185 (2009).
3. D. A. B. Miller, “Attojoule Optoelectronics for Low-Energy Information Processing and Communications,” J. Lightwave. Tech. **35** (3), 343-393 (2017).
4. J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, 1968).
5. 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,” Applied Optics **26** (21), 4639-4652 (1987).
6. P. Ambs, “Optical Computing: A 60-Year Adventure,” Adv. Opt. Tech. **2010,** Article ID 372652, (2010).
7. H. Rajbenbach, Y. Fainman, and S. H. Lee, “Optical implementation of an iterative algorithm for matrix inversion,” Applied Optics, **26** (6), 1024–1031 (1987).
8. K. Wu, C. Soci, P. P. Shum, and N. I. Zheludev, “Computing matrix inversion with optical networks,” Opt. Lett. **22** (1), 295-304 (2014).
9. D. Petrov, Y. Shkuratov, and G. Videen, “Optimized matrix inversion technique for the T-matrix method,” Opt. Lett. **32** (9), 1168–1170 (2007).
10. D. Koller, and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques, Adaptive Computation and Machine Learning Series* (The MIT Press, 2009).
11. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, 2014).
12. Wainwright, M, J. & Jordan, M. I. *Graphical Models, Exponential Families, and Variational Inference* (Now Publishers, 2008).
13. C. Bishop*, Pattern Recognition and Machine Learning* (Verlag New York, 2006).
14. C. Sinoquet and R. Mourad, *Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics* (Oxford Univ. Press, 2014).
15. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice–Hall, Upper Saddle River, 1995).
16. Z. Ghahramani, “Probabilistic machine learning and artificial intelligence,” Nature **521**, 452–459 (2015).
17. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature **521**, 436–444 (2015).
18. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. v. d. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J.

Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, “Mastering the game of go with deep neural networks and tree search,” Nature **529**, 484–489 (2016).

19. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” Nature **518**, 529–533 (2015).
20. S. Z. Li, *Markov Random Field Modeling in Image Analysis* (Springer Science & Business Media, 2009).
21. M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, ‘Wavelet-based statistical signal processing using hidden Markov models,” IEEE Transactions on signal processing, **46** (4), 886-902 (1998).
22. T. Wierschin, K. Wang, M. Welter, S. Waack, and M. Stanke, “Combining features in a graphical model to predict protein binding sites,” Proteins: Structure, Function, and Bioinformatics, **83** (5), 844-852 (2015).
23. G. S. Atsalakis, E. E. Protopapadakis, and K. P. Valavanis, “Stock trend forecasting in turbulent market periods using neuro-fuzzy systems,” Operational Research, **16** (2), 245-269 (2016).
24. A. J. B. Chaney, D. M. Blei, and T. Eliassi-Rad, “A probabilistic model for using social networks in personalized item recommendation,” Proceedings of the 9th ACM Conference on Recommender Systems. ACM, 43-50 (2015).
25. P. A. Blanche et al., "All-optical graphical models for probabilistic inference," 2016 IEEE Photonics Society Summer Topical Meeting Series (SUM), Newport Beach, CA, 2016, pp. 199-200. doi: 10.1109/PHOSST.2016.7548802
26. M. Babaeian, et.al. “Nonlinear optical components for all-optical probabilistic graphical model,” Nature. Commun **9,** 2128 (2018).
27. P-. A. Blanche, M. Babaeian, M. Glick, J. Wissinger, R. Thamvichai, R. A. Norwood, N. Peyghambarian, and M. A. Neifeld, “Optical implementation of probabilistic graphical models,” IEEE. Int. Conf. on Rebooting Computing, (2016).
28. W. E. Moerner, Persistent Spectral Hole-Burning: Science and Applications (Springer-Verlag, Berlin, 1988).
29. M. A. Neifeld, W. R. Babbitt, R. K. Mohan, and A. E. Craig, “Power budget analysis of image-plane storage in spectral hole-burning materials,” J. of Luminescence, **107**, issue 1-4, 114-121 (2004).
30. Y. Jiang, P. T. S. DeVore, and B. Jalali, “Analog optical computing primitives in silicon photonics,” Opt. Lett. **41** (6), 1273–1276 (2016).
31. E. W. V. Stryland, H. Vanherzeele, M. A. Woodall, M. J. Soileau, A. L. Smirl, S. Guha, and T. F. Boggess, “Two photon absorption, nonlinear refraction, and optical limiting in semiconductors,” Opt. Eng. **24**, 613 (1985)
32. M. Hercher, “An analysis of saturable absorbers,” Applied Optics, **6** (5), 947-954 (1967).
33. Cheung, S., Su, T., Okamoto, K. & Yoo, S. J. B. Ultra-compact silicon photonic 512 × 512 25 GHz arrayed waveguide grating router. IEEE J. Sel. Top. Quantum Elect. 20, Issue: 4 (2014).

34. Kamei, S., Ishii, M., Kitagawa, I., Itoh, M. & Hibino, Y. Very low crosstalk arrayed-waveguide grating multi/demultiplexer using cascade connection technique. IEEE Electronics Letters. 36, 823 - 824 (2000).

35. Kamei, S., Ishii, M., Kitagawa, I., Itoh, M. & Hibino, Y. 64-channel ultra-low crosstalk arrayed-waveguide grating multi/demultiplexer module using cascade connection technique. IEEE Electronics Letters. 39, 81 - 82 (2003).

36. Wang, Z. et al. A III-V-on-Si ultra-dense comb laser. Light: Science & Applications. http://dx.doi.org/10.1038/lsa.2016.260 (2017).

37. Klenner, A. et al. Gigahertz frequency comb offset stabilization based on supercontinuum generation in silicon nitride waveguides. Opt. Express 24 (10), 11043-11053 (2016).

38. Ozdur, I. et al. Semiconductor based optical frequency comb source with optical linewidth ≪1 kHz. IEEE. LEOS Conf. Proc. 491-492 (2009).

39. Diddams, S. A., Hollberg, L. & Mbele, V. Molecular fingerprinting with the resolved modes of a femtosecond laser frequency comb. Nature 445, 627-630 (2007).

40. Bartels, A., Oates, C. W., Hollberg, L. & Diddams, S. A. Stabilization of femtosecond laser frequency combs with subhertz residual linewidths. Opt. Letter 29(10), 1081-1083 (2004).

41. Shirasaki, M. Large angular dispersion by a virtually imaged phased array and its application to a wavelength demultiplexer. Opt. Letter 21(5), 366-368 (1996).