

Time wavelength interleaving perceptron at 12 Giga-Ops/s with a Kerr soliton crystal microcomb for optical neural networks

M. Tan¹, X. Xu², and D. J. Moss¹

¹Optical Sciences Centre, Swinburne University of Technology, Hawthorn, VIC 3122, Australia

²Dept. of Electrical and Computer Systems Engineering, Monash University, Clayton, 3800 VIC, Australia

Keywords—Optical neural networks, microwave photonics, micro-ring resonators, Kerr micro-combs.

Optical artificial neural networks (ONNs) have significant potential for ultra-high computing speed and energy efficiency. We report a novel approach to ONNs that uses integrated Kerr optical micro-combs. This approach is programmable and scalable and is capable of reaching ultra-high speeds. We demonstrate the basic building block ONNs — a single neuron perceptron — by mapping synapses onto 49 wavelengths to achieve an operating speed of 11.9×10^9 operations per second, or Giga-OPS, at 8 bits per operation, which equates to 95.2 gigabits/s (Gbps). We test the perceptron on handwritten-digit recognition and cancer-cell detection — achieving over 90% and 85% accuracy, respectively. By scaling the perceptron to a deep learning network using off-the-shelf telecom technology we can achieve high throughput operation for matrix multiplication for real-time massive data processing.

I. INTRODUCTION

Artificial neural networks (ANNs) have achieved significant success in making predictions and achieving simple representations of complex and high dimension data. When sufficient data are used for training, ANNs can outperform computational algorithms [1-5] and even humans for many tasks ranging from the recognition of images to translation of languages, risk assessment and intriguingly, complex board games [6]. The speed and computational power of ANNs is determined by matrix multiplication operations, or multiply-and-accumulate (MAC) operations. Electronic ANN chips include the IBM TrueNorth and Google TPU chips [7, 8]. They use extremely large-scale processor arrays that include the systolic array [8], to enhance the parallelism to achieve operational speeds greater than 180×10^{12} operations per second (Tera-OPS). However, in spite of this performance, since they are electronically based they are still subject to relatively inefficient digital protocols and bandwidth bottlenecks due to the von Neuman effect [9]. In fact, each individual processor is limited in speed to only about 700 MHz [10 9].

Photonic approaches towards ANNs, or optical neural networks (ONNs), are next generation neuromorphic processors and are attracting extremely high levels of interest currently [11-17]. They are highly promising since they offer the potential to achieve extremely high processing speeds [3]. The key is to achieve the weighted synapses that connect the nodes and neurons. In contrast to electronic digital systems that store the synapses in memory, photonic systems operate by realizing actual physical embodiments of synapses, where their number determines the scale of the network, and which depend on the physical parallelism that is fundamentally analog in nature.

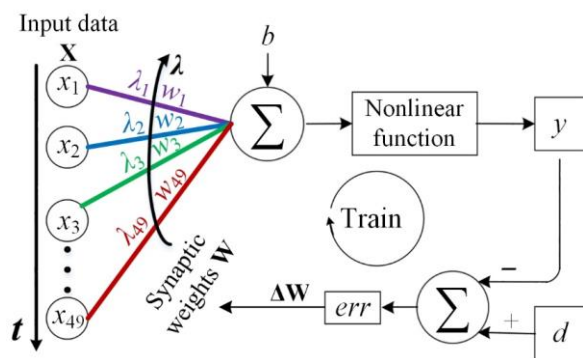


Figure 1. Mathematical model of perceptron.

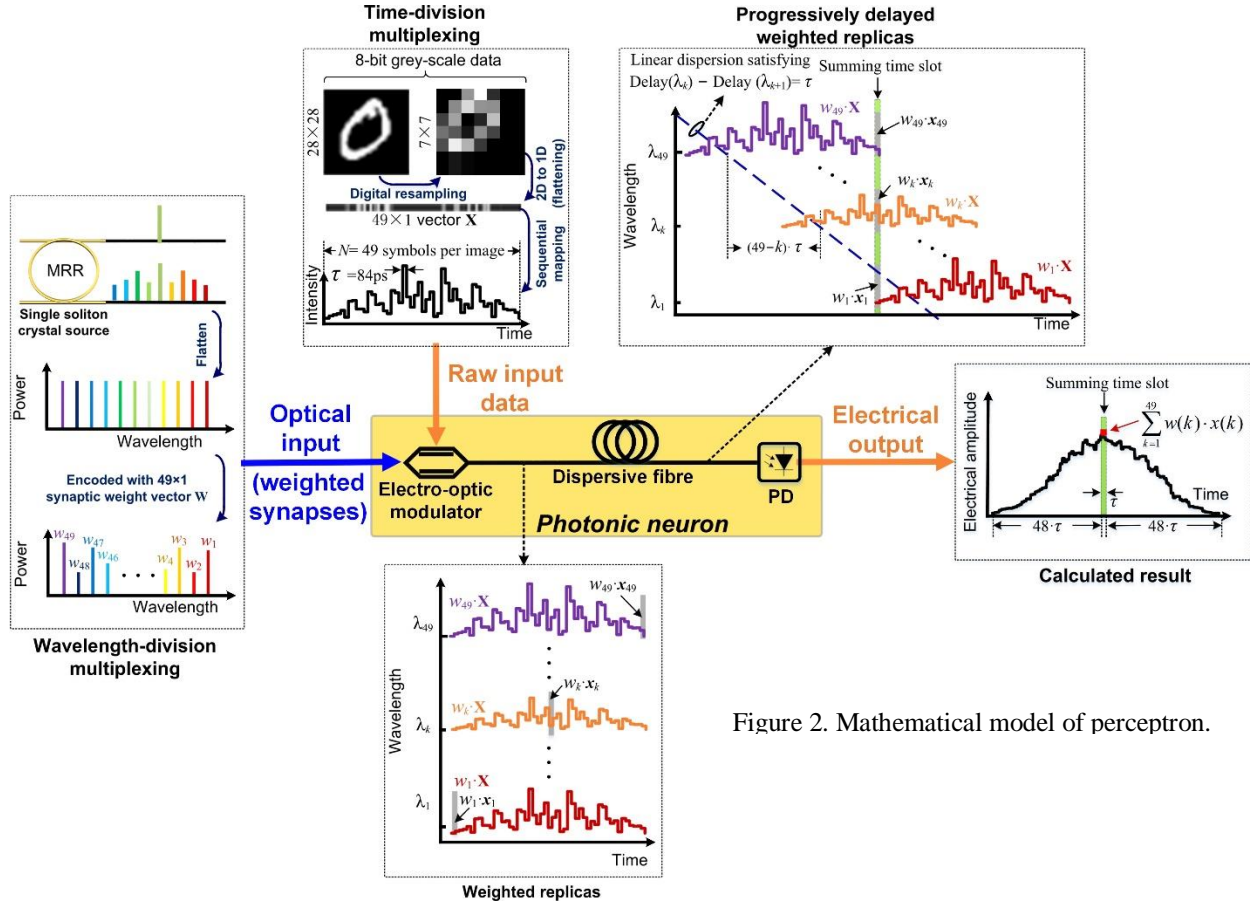


Figure 2. Mathematical model of perceptron.

ONNs have achieved substantial success by using a number of approaches to multiplex synapses in parallel. Schemes based on spatial multiplexing include coherent integrated photonic chips [3] and diffractive bulk optics [17]. These have successfully achieved classification of alphabetic-numeric characters including handwritten digits and vowels. Furthermore, they have achieved low power levels, although they have tradeoffs between the system footprint and processing power, determined by the number of synapses or degree of parallelism. There are a number of ways to realise ONNs, including reservoir computing [18-21] as well as spike processing [22-25]. These both use sophisticated schemes to multiplex the synapses and are both very compact. Photonic reservoir computing multiplexes the synapses in time to achieve very large scale systems with hundreds of input layer nodes. Conversely, spike processors have successfully achieved pattern recognition by using phase change materials in integrated devices [24]. This approach operates via wavelength division multiplexing (WDM) and also benefits from a dynamically reconfigurable operation bandwidth [17-19]. Despite the success of these approaches, they still face limitations. Temporal multiplexing is challenging to dynamically train and to scale up to deep learning systems with multiple layers. Spike processing is limited in the degree of parallelism it can achieve because it relies on arrays of discrete laser diodes. The combined use of temporal, wavelength, and spatial multiplexing has the greatest potential to achieve the highest combination of processing power, operation speed and scale of the network, and this is what our approach uses.

II. PERCEPTRON

Here, we propose [11, 12] a novel scheme for ONNs based on integrated micro-combs to achieve simultaneous temporal, spatial, and wavelength multiplexing, which we then use to perform the dot product of vectors. We perform matrix operations by first flattening the matrices to convert them into vectors at high data rates. Our system is capable of dynamic training and its network structure is highly scalable. We demonstrate a single photonic neuron perceptron with 49 synapses, or wavelengths using the microcomb. Our fundamental building block for ONNs achieves a speed for matrix multiplication at 11.9 billion (Giga) operations/s (OPS) – or GOPs - that equates to 95.2 Gigabits/s for 8 bit operations. We do this via simultaneous synapse weighting in the wavelength domain and the temporal domain, scaling the input data. The

Figure 1 depicts the neuron perceptron mathematical model [25] and Figure 2 outlines the experiment setup that uses a Kerr optical micro-comb source. The perceptron is based on wavelength multiplexing of 49 microcomb wavelengths, done simultaneously with temporal multiplexing, in order to form a single synapse. The main operation consists of matrix multiplication with vectors formed from flattened matrices. The matrix multiplication occurs between the electronic image input data and the synaptic weights, and this is performed with multiple steps using photonics. The input data for classification consists of 28×28 electronic digital matrices with 8-bit grey-scale intensity resolution, which is initially down sampled digitally into 7×7 matrices that are then reorganized into 1D vectors: $X(i) = [X(1), X(2) \dots X(49)]$, which are then multiplexed sequentially in the temporal domain by an electronic high speed D/A converter at 11.9 Gigabaud. Here, each symbol corresponds to the 8-bit pixel input data images and takes up one time slot 84 ps in length. Hence, the whole duration of the waveform is $N \times \tau = 4.12$ ns with $N=49$. In approaches based on digital electronics, the neural network input nodes usually reside in electrical memory and are routed according to memory address. By comparison, the input nodes for the ONN are temporally defined by multiplexing the symbols that are then routed according to their location in time.

$$\mathbf{X} \times \mathbf{W}^T = \begin{pmatrix} w(1) \cdot x(1) & w(1) \cdot x(2) & w(1) \cdot x(3) & \cdots & w(1) \cdot x(49) \\ w(2) \cdot x(1) & w(2) \cdot x(2) & w(2) \cdot x(3) & \cdots & w(2) \cdot x(49) \\ w(3) \cdot x(1) & w(3) \cdot x(2) & w(3) \cdot x(3) & \cdots & w(3) \cdot x(49) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w(49) \cdot x(1) & w(49) \cdot x(2) & w(49) \cdot x(3) & \cdots & w(49) \cdot x(49) \end{pmatrix} \quad (1)$$
$$\begin{pmatrix} & & & w(1) \cdot x(1) & \cdots & w(1) \cdot x(47) & w(1) \cdot x(48) & w(1) \cdot x(49) \\ & & & w(2) \cdot x(1) & w(2) \cdot x(2) & \cdots & w(2) \cdot x(48) & w(2) \cdot x(49) \\ & & w(3) \cdot x(1) & w(3) \cdot x(2) & w(3) \cdot x(3) & \cdots & w(3) \cdot x(49) \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \\ w(49) \cdot x(1) & \cdots & w(49) \cdot x(47) & w(49) \cdot x(48) & w(49) \cdot x(49) \end{pmatrix} \quad (2)$$

Finally, the intensity of all of the optical signals in each time bin are summed via sampling and detection to produce the resulting matrix multiplication (equivalent to a dot product of 49×1 vectors for the case of 7×7 matrices) of the neuron.

given by:

$$\mathbf{X} \cdot \mathbf{W} = \sum_{k=1}^{49} w(k) \cdot x(k) \quad (3)$$

After matrix multiplication, the summed, weighted output is then modulated in order to map it into a desired range by using a nonlinear sigmoid function. In this initial demonstration we achieve this last function using digital electronics, which generates the output of the single neuron perceptron. In principle, however, this can easily be achieved all-optically. Finally, the input data prediction category is produced through comparison between the decision boundary with the neuron output. The decision boundary is a 49 dimensional hyperplane, generated during digital learning carried out offline prior to the experiments. Thus, the input data can be separated into two categories.

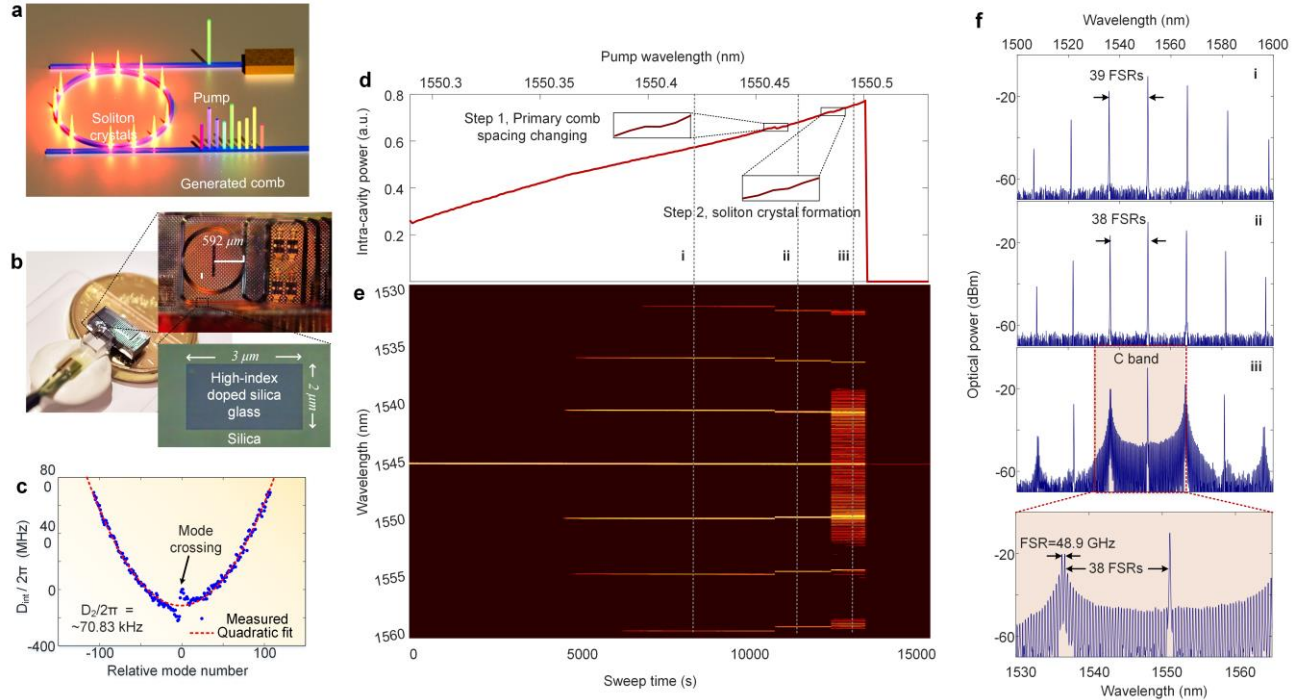


Figure 3 | **a.** Schematic diagram of the soliton crystal microcomb, generated by pumping an on-chip high-Q nonlinear micro-ring resonator with a continuous-wave laser. **b.** Image of the MRR and a scanning electron microscope image of the MRR's waveguide cross section. **c.** Measured dispersion of the MRR showing the mode crossing at ~1552 nm. **d.** Measured soliton crystal step of the intra-cavity power, and **e.** optical spectrum of the microcomb when sweeping the pump wavelength. **f.** Optical spectrum of the generated coherent microcomb at different pump detunings at a fixed power.

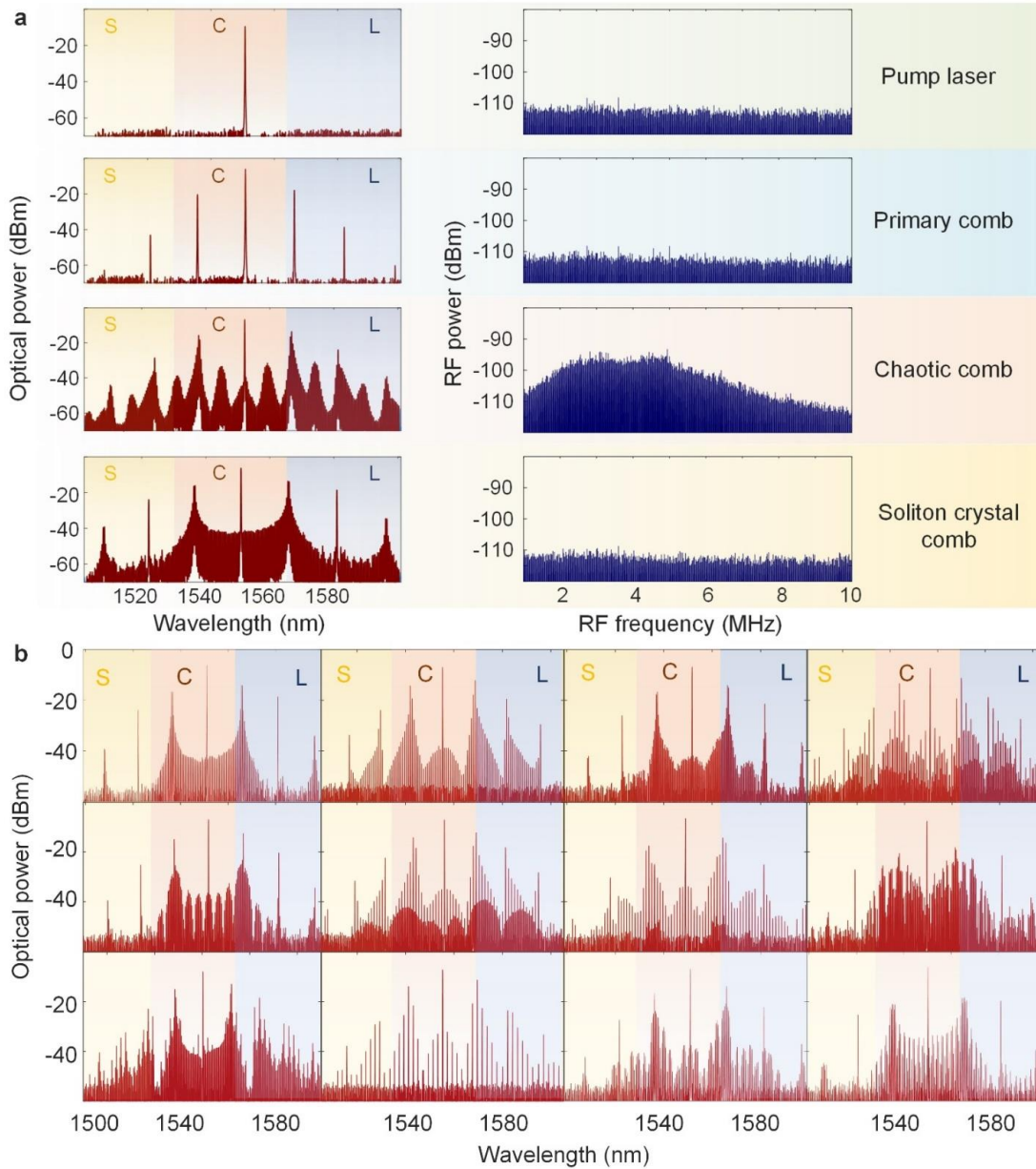


Figure 4 | **a**, Different states and measured RF intensity noise of the microcomb. **b**, Measured low intensity noise states.

III. MICRO-COMBS BASED ON SOLITON CRYSTALS

Kerr optical micro-combs [26-33] have achieved many ground-breaking breakthroughs including optical frequency synthesis [29], ultra-high bit rate data transmission [30], generation of advanced quantum states [31], high level RF signal processing [32], and more. They provide the full capability of mainframe optical frequency combs [34] although in a fully integrated form that has a much more compact footprint as well as the potential to scale the network in power, reliability, and performance [35-42]. The new platforms developed for optical microcombs [27,43,44] have much lower nonlinear absorption than other nonlinear platforms such as chalcogenide glass and semiconductors [45-67].

We use a microcomb that operates via soliton crystal states [68, 69], produced in integrated ring resonators. Soliton crystals display deterministic generation induced by mode crossings that produce a background wave, with all of these processes sustained by the Kerr nonlinearity. For soliton crystals, there is very little nonlinear pump-induced shift in the resonance that otherwise would require difficult dynamic pumping schemes like DKS solitons require [26]. This is because

the intracavity soliton crystal state power is virtually the same as the power for the chaotic state from which it is formed. Hence, very little power jump occurs when they are generated and this allows a reliable and simple method of initiation achievable by simple adiabatic, even manual, tuning of the pump wavelength [68]. This same effect also yields a much higher energy conversion efficiency from pump to comb-line [63]. Soliton crystals have demonstrated a multitude of RF or microwave signal processing based on photonics [11,12,32,70-94]. The integrated ring resonators were made from Hydex glass, a platform that is CMOS compatible [27] (Figure 3). They had a high Q factor of 1.5 million and a 48.9 GHz FSR with a chip to fibre coupling loss of 0.5 dB / facet achieved by on-chip mode converters. The waveguide cross-section of $3\mu\text{m} \times 2\mu\text{m}$ produced anomalous dispersion with a mode crossing near 1552 nm. A 30dBm CW pump laser generated the soliton crystals when its wavelength was swept manually from short to long wavelengths (blue to red) near a resonance.

To generate coherent micro-combs, a CW pump laser (Yenista Tunics – 100S-HP) was employed, with the power amplified to 30dBm by an optical amplifier (Pritel PMFA-37) and the wavelength subsequently swept from blue to red. The acquired soliton crystals optical spectra are shown in Figures 3 and 4. We note that when locking the pump wavelength to the resonance of the MRR, the stability of the microcomb can be further enhanced that could even serve as frequency standards [29]. Figures 3f and 4a show the progression from the initial onset of primary combs, to chaotic combs that are not modelocked, to finally soliton crystal combs. Also shown in Figure 4b are the range of different soliton crystal states that can be obtained by adjusting the pump offset (to the nearest resonance) as well as the overall pumping wavelength. Figure 3 also indicates that the power jump in transitioning from the chaos state to soliton crystal state is extremely small. This arises because the two states have very similar power levels. This is a key reason for the stability of soliton crystals.

IV. EXPERIMENT

As discussed above, the multicasting of the waveform was achieved via intensity modulation of all of the wavelength channels supplied by the shaped comb lines, simultaneously. Hence, the optical signal at the k th ($k=1, 2, \dots, 49$) channel was $w(k) \cdot \mathbf{X}$. The delay that we used for the optical signals at all wavelength channels was 13-km of dispersive single mode fibre which generated a time delay of $(49 - k) \times \tau$ for the k th channel, and τ was measured to be 84 ps. Thus, the optical signals were progressively shifted in the time domain. The optical signal after the single mode fibre was converted to the electrical domain by a photodetector (Finisar VPDV2120), and the waveform was then measured by a high-speed oscilloscope (Keysight DSOZ504A). The sampled output of the photodetector was added to the bias symbol and rescaled in intensity by the reference symbol to extract the recovered output of the ONN and locate the hyper-plane (a trained subspace in the high-dimension space of the input data, which serves as a decision boundary that separates different classes of data).

During the experiment, the 7×7 gray scale data of the handwritten digit figures were first converted into a one dimension array $\mathbf{X}=[x(1), x(2), \dots, x(49)]$ by assembling each column head-to-tail. Then a 49-symbol waveform was generated and coded with the intensities at each time slot in proportion to the values of \mathbf{X} at corresponding sequences, thus the input data \mathbf{X} were multiplexed in the time domain. The 49-symbol waveform was generated by an arbitrary waveform generator (Keysight M8195A), which supported a sample rate of 65 Giga-Samples/s and an analog bandwidth of up to 25 GHz. To acquire stepwise waveforms for the input nodes, we used 5 sample points at 59.421642 Giga Samples/s to form each single symbol of the input waveform, which also matched the progressive time delay τ (84 ps) of the dispersive buffer.

The optical power of the 49 microcomb lines was shaped according to the intensity of pre-trained neuron weights $\mathbf{W}=[w(1), w(2), \dots, w(49)]$. We shaped the comb lines' power with a programmable optical spectral shaper using liquid crystal on silicon techniques (Finisar WaveShaper 4000S), which could dynamically reconfigure the ONN connections within 500 ms with a resolution of 1 GHz. Two stages of programmable optical spectral shapers were employed for a larger loss dynamic range. The first WaveShaper was used to flatten the microcomb, while the second one was used to achieve pre-trained neuron weights. A feedback loop was used to enhance the shaping accuracy, where the comb lines' power after shaping was measured by an optical spectrum analyser (Anritsu MS9710C) and compared with the pre-trained weights to generate an error signal for the calibration of the WaveShapers' loss characteristics.

Figure 5 shows the time-domain multiplexed input layer for the cancer diagnosis test. The generated 11.9 Giga-baud data stream of the encoded 75 sets of features shows the 30-symbol encoded data for each set and 3 symbols padded for post measurement, including a trigger symbol to trigger the oscilloscope, a reference symbol to calibrate the reference level, and a bias symbol encoded with the pre-trained bias to locate the decision boundary. Figure 6 shows the experimental recognition of cancer diagnosis. Figure 6a shows the optical spectrum of the shaped (soliton crystal) micro-comb measured by an optical spectrum analyser, while Figure 6b shows the measured and sampled output waveform from the

photodetector. Figure 6c shows the recovered ONN predictions $X \times W + b$ acquired by rescaling the sampled results via the reference symbol, and the hyper-plane $X \times W + b = 0$ (black line).

V. DATASETS AND PRE-TRAINING

The datasets we employed was from MNIST (Modified National Institute of Standards and Technology) handwritten digit database [95] and part of the publicly available Wisconsin Breast Cancer dataset [96]. The datasets of recognition tasks were first separated as training sets and test sets. The training sets were used for the offline training with the Back Propagation algorithm [97], performed on an electronic computer using Matlab™, to acquire pre-trained weights and bias. The test sets were tested with both the ONN and an electronic computer for comparison. We note that, since the number of training sets is sufficiently large compared with the number of synaptic connections, the cross validation process was not employed in this work—and in any case, it could be performed offline before the pre-training.

We note here that the accuracy of the ONN predictions was experimentally limited by the performance of the arbitrary waveform generator, which introduced errors to the symbols' intensities and thus deteriorated the correctness of the matrix multiplication. This can be addressed by using an arbitrary waveform generator with a larger analog bandwidth, or a higher sampling rate. Addressing this issue would result in higher levels of correctness than reported here.

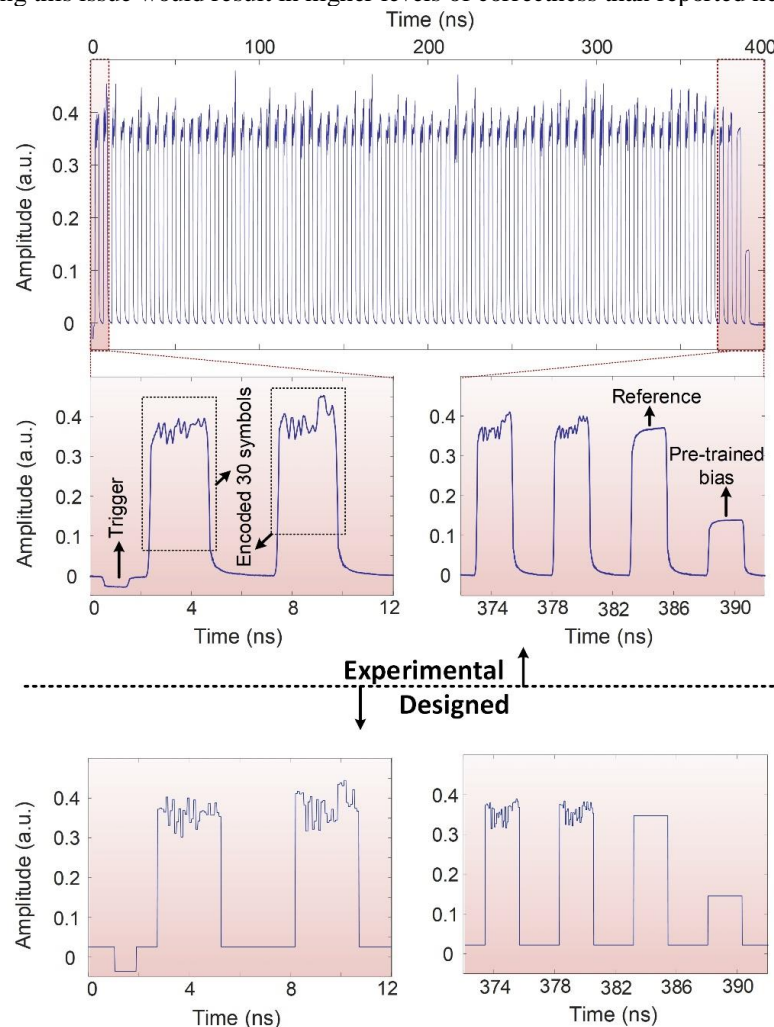


Figure 5 | Time-domain multiplexed input layer of cancer diagnosis test. Generated 11.9 Giga-baud data stream of the encoded 75 sets of features showing 30-symbol encoded data for each set and 3 symbols padded for post measurement, including a trigger symbol to trigger the oscilloscope, a reference symbol to calibrate the reference level, and a bias symbol encoded with the pre-trained bias to locate the decision boundary.

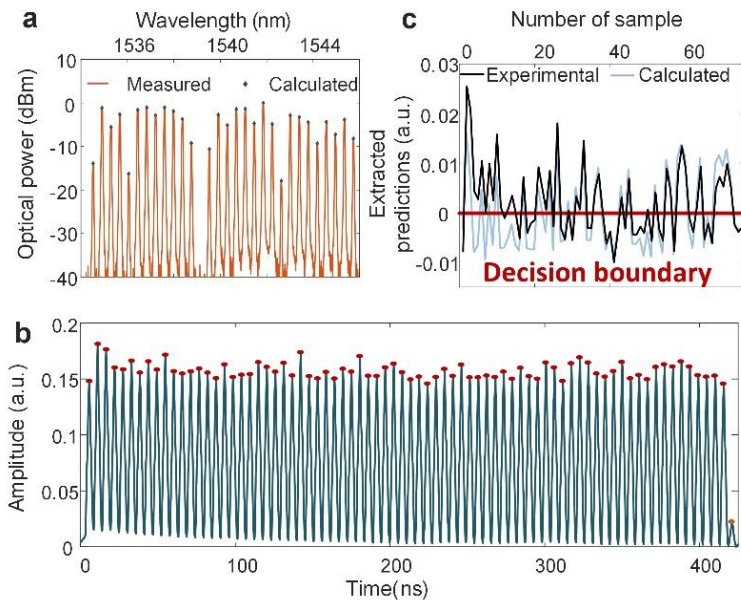


Figure 6 | Experimental recognition of cancer diagnosis. **a**, Optical spectrum of the shaped (soliton crystal) micro-comb measured by an optical spectrum analyser. **b**, Measured and sampled output waveform from the photodetector. **c**, Recovered ONN predictions $X \times W + b$ acquired by rescaling the sampled results via the reference symbol, and the hyper-plane $X \times W + b = 0$ (black line).

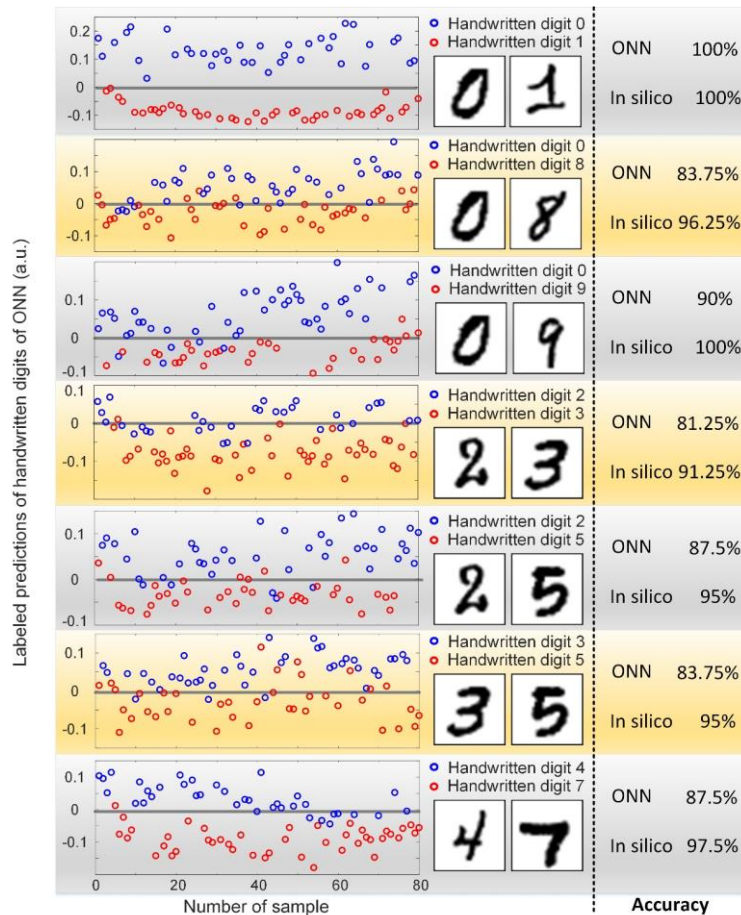


Figure 7 S2 | ONN predictions of handwritten digits labeled according to their correct answers.

VI. RESULTS

First, we evaluated the performance of the network using a number of handwritten digit pairs from a body of 500 images for each digit, from which we randomly selected 920 images for prior off-line training, which left 80 figures to evaluate the system performance. The handwritten digital images were electronically down-sampled to reduce the size of the images to 7×7 from 28×28 . Next, this was transformed into a 49 symbol one dimensional array, following which the array was temporally multiplexed with each symbol occupying an 84ps time slot, yielding a modulation rate of 11.9 Gigabaud. The data vector dimension of our perceptron needed to match the weight vector dimension, given by number of wavelength, which was 49. Therefore, we used a down-sampling method on the image to reduce the length of the vector to 49.

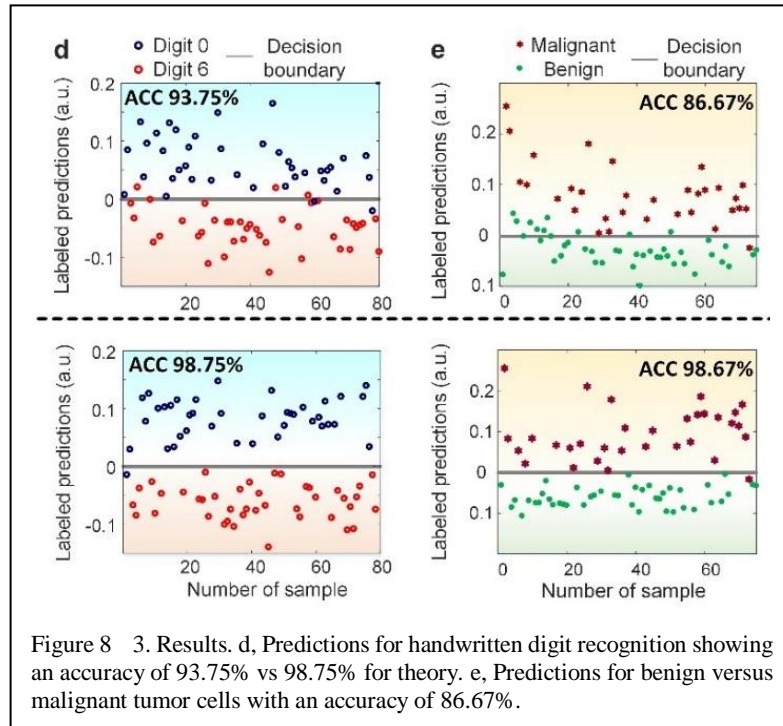


Figure 8 3. Results. d, Predictions for handwritten digit recognition showing an accuracy of 93.75% vs 98.75% for theory. e, Predictions for benign versus malignant tumor cells with an accuracy of 86.67%.

matrix multiplication computations on the multiple input data samples were then compared in intensity against this decision boundary, finally producing the predictions of the perceptron (Figures 7, 8). We tested the perceptron performance for classifying 2 benchmark tests delineated by the decision boundary – first for two handwritten digits (0 and 6), followed by determining whether cancer cells are benign or malignant. For the handwritten digits the perceptron produced an accuracy of 93.75%, versus 98.75% that can be achieved with an electronic digital neural network. For the tissue biopsy data classification for cancer cells (Figure 8), individual cell nuclei were extracted from breast mass tissue via fine needle aspirate and then imaged with a microscope. These images were previously characterized to distinguish 30 different features including texture, perimeter, radius, etc.. For our experiments, the data for 521 cell nuclei were used for pre-training the network, with a further 75 used as the basis for the testing diagnosis. This follows a very similar process to that used for the handwritten digit tests discussed above. We obtained an 86.67% accuracy versus 98.67% that can be achieved with a digital electronic neural network.

In our experiments we used Intel's approach of evaluating digital microprocessors [98]. Since our system is rather more complex in that it uses input data and weight vectors for the MAC calculations that come from different sources that are multiplexed in time and wavelength, we define the throughput speed based on the temporal data sequence of the electronic output port, in order to be unambiguous. According to the protocol of broadcast-and-delay, each computation cycle consists of one vector dot product between the 49 symbol data and the weighted vectors, resulting in a time data sequence having a length of $48+1+48$ symbols, yielding a total duration time of $97 \times 84\text{ps}$. The 49th symbol represents the desired result – i.e., the vector dot product resulting from 49 MAC operations, and hence the perceptron throughput is given by $49 / (84\text{ps} \times 97) = 5.95$ Giga-MACs/s. Since each MAC operation consists of two operations – a multiply followed by an accumulate operation – our throughput measured in operations (OPS) is twice that measured in MACs/s, or $(49 \times 2) / (84\text{ps} \times 97) = 11.9$ Giga-OPS.

The input data sequence contained 8-bit symbols of 256 discrete levels, reflecting the pixel values of the grey scale image. The 8 bits was limited by our electronic arbitrary waveform generator's intensity resolution. The Waveshaper had a range in attenuation of 35 dB, which is equivalent to a resolution of 11 bits or 33 dB ($=10 \times \log_{10}[2^{11}]$). Therefore, every computing cycle had an effective throughput bit rate of $(49 \times 2) \times 8 / (84 \text{ ps} \times 97) = 95.2$ Gigabits/s. For analogue systems such as ours, both the intensity resolution and the bit rate are limited by the system SNR (signal-to-noise ratio). Therefore, in order to have a full resolution of 8-bits, our system needed to have a SNR greater than $20 \cdot \log_{10}(28) = 48$ dB in terms of electric power. This is well within the capability of analogue photonic microwave links, such as the perceptron system that we reported here which had an OSNR > 28 dB.

Our perceptron is the fastest optically based neuromorphic processor ever reported, although making direct comparisons with all of the different approaches is challenging since they vary so widely. As an example, on the one hand systems based on static or continuous sources that perform one-off or single-shot measurements [11, 17, 24] can have a very low latency. However, on the other hand, they also suffer from an extremely low throughput since the input data cannot be in any rapid manner. While our perceptron did have a relatively large latency of $\sim 64 \mu\text{s}$, this was purely due to the dispersive delay component which in our case was a simple spool of optical fibre. This did not, however, have any effect on the speed or throughput of our system. Moreover, in fact this can be dramatically reduced or virtually eliminated – easily to less than 200 ps – just by using any type of compact device that can replace the dispersive delay of the fibre, such as sampled Bragg gratings or etalon based tuneable dispersion compensators [99-103].

VII. SPEED CALCULATION

Following our definition of throughput and latency introduced in the manuscript, the overall throughput of the deep ONN is roughly the product of each hidden layer's speed and the number of hidden layers, although we note that rigorous and accurate calculation of the throughput is only possible with specific configurations of the network.

Here is a simple example of calculation (this example is just to show the calculations of throughput and latency, the actual performance in terms of prediction accuracies is not the focus of our discussion here): the input waveform/layer is the same as the demonstrated perceptron (49×1 vector at 11.9 Giga Baud with 8-bit resolution, $\tau = 84\text{ps}$), the network has a hidden layer that each has 7 fully connected neurons, and an output layer that has 10 fully connected neurons (to match with the number of categories for digits from 0 to 9). As a result, 343 (49×7) and 70 (7×10) wavelengths would be needed in the hidden and output layer, respectively. This can be achieved by using smaller FSR microcombs such as 25GHz across the wide optical band (the C + L bands already reach >11THz wide).

In the hidden layer, each initial electrical output waveform (right after the photodetection and before the digital signal processing) corresponds to the output of a single neuron and has a duration of $(49 \times 2 - 1) \times 84\text{ps} = 8.148 \text{ ns}$. Only one time slot of each group of symbols represents the result of matrix multiplication between the input vector and the weight vector that constitutes of $49 \times 2 = 98$ floating point operations. As a result, the throughput of each neuron is given as $98 / 8.148 = 12.0275$ Giga-OPS. Since different neurons are multiplexed in both the spatial and wavelength domain and detected in parallel, the total throughput of the hidden layer would be $12.0275 \times 7 = 84.1925$ Giga-OPS.

In the output layer, the generated electrical waveform of each neuron has a duration of $(7 \times 2 - 1) \times 84\text{ps} = 1.092 \text{ ns}$. Only one time slot of each group of symbols represents the result of matrix multiplication between the input vector (sampled and re-multiplexed waveform from the hidden layer) and the weight vector that constitutes of $7 \times 2 = 14$ floating point operations, thus the throughput would be $14 / 1.092 = 12.8205$ Giga-OPS for each neuron and the total throughput of output layer would be $12.8205 \times 10 = 128.205$ Giga-OPS. As such, the total peak throughput of the network would be $84.1925 + 128.205 = 212.3975$ Giga-OPS. In addition, the latency of the overall network is the sum of each layer's latency, which mainly comes from the dispersive optical buffer and the electrical sampling and multiplexing module. We assume the latency to be 200 ps for the buffer in integrated forms and to be twice of the waveform duration for the re-sampling unit ($2 \times 8.148 \text{ ns}$ and $2 \times 1.092 \text{ ns}$ for the hidden and output layer, respectively), the total latency of the example network would roughly be 18.68 ns. We note that the latency is just a very rough estimation showing how to calculate or measure the performance of our approach, the practical calculations of the latency are subject to more detailed parameters.

The speed of the network has the potential to reach 10 Tera-OPS [12], determined as follows. With 20 layers, each layer featuring 20 neurons and a modulation rate of 25 Giga baud, the overall throughput should be around $20 \times 20 \times 25 = 10$ tera-OPS, according to the discussion in the above section. With 8-bit resolution, the total potential throughput in terms of bit rate could reach $10 \times 8 = 80$ Tbps. We note that other widely used techniques in telecommunications such as polarization

multiplexing and coherent modulation formats could also potentially boost the computing speed of the proposed neuron network in this work.

Table 1 shows the performance matrices of state-of-art ONNs. We note that it is difficult to directly compare different kinds of ONNs, since on one hand, there are no universal and specific definitions of ONN's parameters. On the other hand, the operation principles of existing ONNs are quite different and have their unique advantages. As such, here we highlight the decent advances of existing works and focus on the speed parameters, including the *latency* and *throughput*, to reflect our ONN's advantages in this aspect.

Table 1 Performance comparison of state-of-the-art ONNs

Approach\Parameter		Compatibility with digital electronics	Latency	Throughput speed per unit	
				OPS	bits/s
Diffraction devices [17]		—	< 10 ns	—	—
Integrated couplers [3]		—	< 0.1 ns	—	—
Reservoir computing [20]		Yes	< 1 μs	17.6 G	—
Spike computing [23]		Yes	< 1 μs	8 G	8 G
Spike computing [24]		—	< 0.1 μs	—	—
[11]	Single Perceptron	Yes	64 μs	11.9 G	95.2 G
[12]	Deep ONN	Yes	>18.68 ns	>10 T	>80 T

“—” denotes the corresponding parameter is either not demonstrated or not indicated in the work.

VIII. CONCLUSIONS

We report an optical neural network consisting of a single perceptron that operates with an integrated optical Kerr micro-comb source. The system achieves a single processor throughput speed of 11.9 Giga-OPS/s, equivalent to 95.2 Gigabits/s. We demonstrate benchmark tests including cancer cell diagnosis and handwritten digit recognition. We outline different approaches to scale the network to deep learning ONN architectures that have significantly increased processing power and throughput speed. This is possible because of the high level of parallelism that can be realized via simultaneous time, spatial, and wavelength multiplexing. Our results are a consequence of the superior stability, reliability and efficiency of optical Kerr soliton crystal microcombs [104-106]. Our approach has significant possibilities for real-time analysis of high dimensional data for advanced applications.

REFERENCES

- [1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436 (2015).
- [2] R. J. Schalkoff, *Pattern recognition*. Wiley Encyclopedia of Computer Science and Engineering (2007).
- [3] Y. Shen *et al.*, “Deep learning with coherent nanophotonic circuits”. *Nature Photonics* **11**, 441 (2017).
- [4] X. Xu *et al.*, “Photonic perceptron based on a Kerr microcomb for scalable high speed optical neural networks”, *Laser and Photonics Reviews*, vol.14, no. 8, 2000070 (2020). DOI:10.1002/lpor.202000070.
- [5] V. Mnih *et al.*, “Human-level control through deep reinforcement learning”. *Nature* **518**, 529 (2015).
- [6] D. Silver *et al.*, “Mastering the game of Go without human knowledge”. *Nature* **550**, 354 (2017).
- [7] S. Esser *et al.*, “Convolutional networks for fast, energy-efficient neuromorphic computing”. *Proceedings of the National Academy of Sciences* **113**, 11441, doi:10.1073/pnas.1604850113 (2016).
- [8] A. Graves *et al.*, “Hybrid computing using a neural network with dynamic external memory”. *Nature* **538**, 471–476 (2016).
- [9] Ambrogio, S. *et al.*, “Equivalent-accuracy accelerated neural-network training using analogue memory”, *Nature* **558**, 60–67 (2018).
- [10] D. A. B. Miller, “Attojoule Optoelectronics for Low-Energy Information Processing and Communications”. *Journal of Lightwave Technology* **35**, 346-396 (2017).
- [11] X. Xu, M. Tan, J. Wu, S. T. Chu, B. E. Little, R. Morandotti, A. Mitchell, B. Corcoran, D. Hicks, and D. J. Moss, “Photonic perceptron based on a Kerr microcomb for scalable high speed optical neural networks”, IEEE Topical Meeting on Microwave Photonics (MPW), pp. 220-224, Matsue, Japan, November 24-26, 2020. Electronic ISBN: 978-4-88552-331-1.
- [12] X. Xu, *et al.*, “11 TOPs photonic convolutional accelerator for optical neural networks”, *Nature* **589**, 44-51 (2021).

- [13] B. J. Shastri *et al.*, “Photonics for artificial intelligence and neuromorphic computing”, *Nature Photonics* **15**, (2) 102-114 (2021).
- [14] H. Wu, and Q. Dai, “Artificial intelligence accelerated by light”, *Nature* **589**, 25-26 (2021).
- [15] Feldmann, J. *et al.*, “Parallel convolutional processing using an integrated photonic tensor core”, *Nature* **589**, 52-58 (2021).
- [16] G Wetzstein *et al.*, “Inference in artificial intelligence with deep optics and photonics”, *Nature* **588** (7836), 39-47 (2020).
- [17] X. Lin *et al.*, “All-optical machine learning using diffractive deep neural networks”. *Science* **361**, 1004-1008 (2018).
- [18] P. Antonik *et al.*, “Human action recognition with a large-scale brain-inspired photonic computer,” *Nature Machine Intelligence*, **1**, 530-537 (2019).
- [19] L. Appeltant *et al.*, “Information processing using a single dynamical node as complex system. *Nature Communications* **2**, 468 (2011).
- [20] L. Larger *et al.*, “High-speed photonic reservoir computing using a time-delay-based architecture: Million words per second classification”. *Physical Review X* **7**, 011015 (2017).
- [21] H.-T. Peng *et al.*, “Neuromorphic Photonic Integrated Circuits”. *IEEE Journal of Selected Topics in Quantum Electronics*, **24**, 6101715(2018).
- [22] Tait *et al.*, “Broadcast and Weight: An Integrated Network for Scalable Photonic Spike Processing”. *Journal of Lightwave Technology* **32**, 4029-4041 (2014).
- [23] Tait *et al.*, “Demonstration of WDM weighted addition for principal component analysis”. *Optics Express* **23**, 12758-12765 (2015).
- [24] J. Feldmann *et al.*, “All-optical spiking neurosynaptic networks with self-learning capabilities”. *Nature* **569**, 208-214 (2019).
- [25] F. Rosenblatt, *The perceptron: a probabilistic model for information storage and organization in the brain*. *Psychological Review* **65**, 386 (1958).
- [26] T. J. Kippenberg, R. Holzwarth, S. A. Diddams, “Microresonator-based optical frequency combs”. *Science* **332**, 555-559 (2011).
- [27] D. J. Moss, R. Morandotti, A. L. Gaeta, M. Lipson, “New CMOS-compatible platforms based on silicon nitride and Hydex for nonlinear optics”. *Nature photonics*, vol. **7**, no. 8, pp.597–607 (2013).
- [28] A. A. Savchenkov *et al.*, “Tunable optical frequency comb with a crystalline whispering gallery mode resonator”. *Physics Review Letters* **101**, 093902 (2008).
- [29] D. T. Spencer *et al.*, “An optical-frequency synthesizer using integrated photonics”. *Nature* **557**, 81-85 (2018).
- [30] P. Marin-Palomo *et al.*, “Microresonator-based solitons for massively parallel coherent optical communications. *Nature*, vol. **546**, 274 (2017).
- [31] M. Kues *et al.*, “Quantum optical microcombs”. *Nature Photonics* vol. **13**, no. 3, 170-179 (2019).
- [32] X. Xu, M. Tan, J. Wu, R. Morandotti, A. Mitchell and D. J. Moss, “Microcomb-based photonic RF signal processing”. *IEEE Photonics Technology Letters*, vol. **31**, no. 23, 1854-1857 (2019).
- [33] A. Pasquazi *et al.*, “Micro-combs: a novel generation of optical sources”. *Physics Reports*, vol. **729**, 1-81 (2018).
- [34] J. L. Hall, Nobel Lecture: Defining and measuring optical frequencies. *Reviews of Modern Physics* **78**, 1279 (2006).
- [35] M. Ferrera *et al.*, “Low-power continuous-wave nonlinear optics in doped silica glass integrated waveguide structures,” *Nat. Photonics*, vol. 2, no. 12. pp. 737-740 (2008).
- [36] M. Peccianti *et al.*, “Demonstration of a stable ultrafast laser based on a nonlinear microcavity,” *Nat. Commun.*, vol. **3**, 765, pp. 1-6 (2012).
- [37] M. Peccianti *et al.*, “Sub-picosecond optical pulse compression via an integrated nonlinear chirper,” *Optics Express*, vol. **18**, no. 8, pp. 7625-7633, Apr. 2010.
- [38] H. Bao *et al.*, “Type-II micro-comb generation in a filter-driven four wave mixing laser,” *Photonics Research*, vol. **6**, no. 5, pp. B67-B73, May 2018.
- [39] C. Reimer, et al., “High-dimensional one-way quantum processing implemented on d-level cluster states”, *Nature Physics*, vol. 15, no.2, pp. 148–153, 2019.
- [40] A. Pasquazi, et al., “Sub-picosecond phase-sensitive optical pulse characterization on a chip”, *Nature Photonics*, vol. 5, no. 10, pp. 618-623 (2011). DOI: 10.1038/nphoton.2011.199.
- [41] M. Kues, et al., “Passively modelocked laser with an ultra-narrow spectral width”, *Nature Photonics*, vol. 11, no. 3, pp. 159, 2017. DOI:10.1038/nphoton.2016.271
- [42] H. Bao, et al., “Laser cavity-soliton microcombs,” *Nature Photonics*, vol. 13, no. 6, pp. 384-389, Jun. 2019.
- [43] D. Duchesne, M. Peccianti, M. R. E. Lamont et al., “Supercontinuum generation in a high index doped silica glass spiral waveguide,” *Optics Express*, vol. 18, no. 2, 923-930 (2010).
- [44] M. Ferrera, et al., “On-chip CMOS-compatible all-optical integrator”, *Nature Communications*, vol. 1, Article 29 (2010). DOI:10.1038/ncomms1028
- [45] M. Rochette, J. N. Kutz, J. L. Blows, D. Moss, J. T. Mok, and B. J. Eggleton, “Bit Error Ratio Improvement with 2R Optical Regenerators”, *IEEE Photonics Technology Letters*, vol. 17, no. 4, pp. 908-910 (2005).
- [46] A. Pasquazi, Y. Park, J. Azana et al., “Efficient wavelength conversion and net parametric gain via Four Wave Mixing in a high index doped silica waveguide,” *Optics Express*, vol. 18, no. 8, 7634-7641 (2010).
- [47] M. Shokoh-Saremi et al., “High performance Bragg gratings in chalcogenide rib waveguides written with a modified Sagnac interferometer: experiment and modeling”, *Journal of the Optical Society of America B*, vol. 23, 1323 (2006).
- [48] M. Rochette, L.B. Fu, V.G. Ta’eed, I.C.M. Littler, D.J. Moss, B.J. Eggleton, “2R Optical Regeneration: Beyond Noise Compression to BER Reduction”, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 12, 736 (2006).

- [49] M. Rochette et al., "Bit Error Ratio Improvement with 2R Optical Regenerators", IEEE Photonics Technology Letters, vol. 17, no. 4, pp. 908-910 (2005).
- [50] M.R.E. Lamont et al., "Error-free wavelength conversion via cross phase modulation in 5 cm of As₂S₃ chalcogenide glass rib waveguide", Electronics Letters, vol. 43, 945 (2007).
- [51] M. Lee et al., "Photosensitive post tuning of photonic crystal waveguide resonant wavelengths", Optics Express, vol. 15, 1277 (2007).
- [52] C. Grillet et al., "Nanowire coupling to photonic crystal nanocavities for single photon sources", Optics Express, vol. 15, 1267 (2007).
- [53] D. Freeman, et al., "Chalcogenide Glass Photonic Crystal Devices", Photonic and Electromagnetic Crystal Structures, Photonics and Nanostructures-Fundamentals and Applications, Vol. 6, pp. 3-11 (2008). doi:10.1016/j.photonics.2007.11.001.
- [54] T. Monro, D.J.Moss, M. Bazylenko, C. Martijn de Sterke, and L. Poladian, "Observation of self-trapping of light in a self written channel in photosensitive glass", Physical Review Letters, vol. 80, 4072 (1998).
- [55] M.D. Pelusi et al., "High bit rate all-optical signal processing in a fiber photonic wire", Optics Express, vol. 16, pp. 11506-11512 (2008).
- [56] S. Tomljenovic-Hanic, M.J. Steel, C. Martijn de Sterke and D. J. Moss, "Photonic crystal slab hetero-structures formed by refractive index variations in chalcogenide glasses", Optics Letters, vol. 32, 542 (2007).
- [57] D.J.Moss, L.Fu, I.Littler, B.J.Eggleton, "Ultra-high speed all-optical modulation via two-photon absorption in silicon-on-insulator waveguides", Electronics Letters, vol. 41, 320 (2005).
- [58] B. Corcoran et al., "Silicon nanowire based radio-frequency spectrum analyzer", Optics Express, vol. 18, no. 19, pp. 20190-20200 (2010).
- [59] T.Ido, H.Sano, D.J.Moss, S.Tanaka, and A.Takai, "Strained InGaAs/InAlAs MQW electroabsorption modulators with large bandwidth and low driving voltage", IEEE Photonics Technology Letters, vol. 6, 1207 (1994).
- [60] M.R.E. Lamont, M. Rochette, D.J. Moss, B.J. Eggleton, "Two-photon absorption effects on self-phase-modulation-based 2R optical regeneration", IEEE Photonics Technology Letters, vol.18, no. 10, 1185-1187 (2006).
- [61] M Ferrera et al., "All-optical 1st and 2nd order integration on a chip", Optics express, vol. 19, no. 23, pp. 23153-23161 (2011).
- [62] M Ferrera et al., "CMOS compatible integrated all-optical radio frequency spectrum analyzer", Optics Express, vol. 22, no. 18, pp. 21488-21498 (2014).
- [63] A Tuniz, G Brawley, DJ Moss, BJ Eggleton, "Two-photon absorption effects on Raman gain in single mode As₂Se₃ chalcogenide glass fiber", Optics Express, vol. 16, no. 22, pp. 18524-18534 (2008).
- [64] E. Ghahramani, D.J. Moss, J.E. Sipe, "Linear and nonlinear optical properties of (GaAs)_m / (AlAs)_n superlattices", Physical Review B, vol. 43, no. 11, 9269 (1991).
- [65] VG Ta'eed et al., "Error free all optical wavelength conversion in highly nonlinear As-Se chalcogenide glass fiber", Optics Express, vol. 14, no. 22, pp.10371-10376 (2006).
- [66] C Monat et al., "Investigation of phase matching for third-harmonic generation in silicon slow light photonic crystal waveguides using Fourier optics", Optics Express, vol. 18, no. 7, pp. 6831-6840 (2010).
- [67] C Grillet et al., "Characterization and modeling of Fano resonances in chalcogenide photonic crystal membranes", Optics Express, vol. 14, no. 1, pp. 369-376 (2011).
- [68] W. Wang *et al*, Robust soliton crystals in a thermally controlled microresonator. Optics Letters **43**, 2002-2005 (2018).
- [69] D. C. Cole *et al*, "Soliton crystals in Kerr resonators". *Nature Photonics* **11**, 671 (2017).
- [70] X. Xu *et al*, "Reconfigurable broadband microwave photonic intensity differentiator based on an integrated optical frequency comb source," APL Photonics, vol. **2**, no. 9, 096104, Sep. 2017.
- [71] J. Wu *et al*, "RF Photonics: An Optical Microcombs' Perspective," IEEE J. Sel. Top. Quantum Electron., vol. **24**, no. 4, 6101020, Jul-Aug. 2018.
- [72] X. Xu *et al*, "Photonic RF phase-encoded signal generation with a microcomb source," Journal of Lightwave Technology, vol. 38, no. 7, pp. 1722-1727, April 1. 2020.
- [73] M. Tan *et al*, "Microwave and RF Photonic Fractional Hilbert Transformer Based on a 50 GHz Kerr Micro-Comb," Journal of Lightwave Technology, vol. **37**, no. 24, pp. 6097-6104. 2019.
- [74] X. Xu *et al*, "Micro-comb based photonic local oscillator for broadband microwave frequency conversion," Journal of Lightwave Technology, vol. **38**, no. 2, pp. 332-338. 2020.
- [75] X. Xu *et al*, "Photonic microwave true time delays for phased array antennas using a 49 GHz FSR integrated optical micro-comb source," Photonics Res., vol. **6**, no. 5, pp. B30-B36, 2018.
- [76] X. Xue *et al*, "Microcomb-based true-time-delay network for microwave beamforming with arbitrary beam pattern control," Journal of Lightwave Technology, vol. **36**, no. 12, pp. 2312-2321, Jun. 2018.
- [77] X. Xu *et al*, "Advanced RF and microwave functions based on an integrated optical frequency comb source," Opt. Express, vol. **26**, no. 3, pp. 2569-2583, Feb. 2018.
- [78] X. Xue *et al*, "Programmable single-bandpass photonic RF filter based on a Kerr comb from a microring," Journal of Lightwave Technology, vol. **32**, no. 20, pp. 3557-3565, Oct. 2014.
- [79] X. Xu *et al*, "Advanced adaptive photonic RF filters with 80 taps based on an integrated optical micro-comb source," Journal of Lightwave Technology, vol. **37**, no. 4, pp. 1288-1295, 2019.

- [80] T. G. Nguyen *et al*, "Integrated frequency comb source-based Hilbert transformer for wideband microwave photonic phase analysis," *Opt. Express*, vol. **23**, no. 17, pp. 22087-22097, Aug. 2015.
- [81] X. Xu *et al*, "Broadband RF channelizer based on an integrated optical frequency Kerr comb source," *Journal of Lightwave Technology*, vol. **36**, no. 19, pp. 4519-4526, 2018.
- [82] X. Xu *et al*, "High performance RF filters via bandwidth scaling with Kerr micro-combs," *APL Photonics*, vol. **4**, no. 2, pp. 026102, 2019.
- [83] M. Tan *et al*, "RF and microwave fractional differentiator based on photonics," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 67, No. 11, pp. 2767-2771. 2020. DOI: 10.1109/TCSII.2020.2965158
- [84] B. Corcoran *et al*, "Ultra-dense optical data transmission over standard fiber with a single chip source", *Nature Communications* vol. 11, Article: 2568, May 22 (2020). DOI:10.1038/s41467-020-16265-x
- [85] T. G. Nguyen *et al*, "Integrated frequency comb source based Hilbert transformer for wideband microwave photonic phase analysis", *Optics Express*, vol. 23, no. 17, 22087 (2015). DOI:10.1364/OE.23.022087
- [86] X. Xu *et al*, "High performance RF filters via bandwidth scaling with Kerr micro-combs," *APL Photonics*, vol. 4, no. 2, pp. 026102, 2019.
- [87] M. Tan *et al*, "Photonic RF and microwave filters based on 49GHz and 200GHz Kerr microcombs", *Optics Communications*, vol. 465, Article: 125563. 2020. DOI:10.1016/j.optcom.2020.125563.
- [88] X. Xu *et al*, "Photonic RF and microwave integrator with soliton crystal microcombs", *IEEE Transactions on Circuits and Systems: Express Briefs*, Vol. 67, Issue: 12, pp. 3582-3586 (2020). DOI:10.1109/TCSII.2020.2995682
- [89] X. Xu *et al*, "Broadband photonic radio frequency channelizer with 90 channels based on a soliton crystal microcomb", *Journal of Lightwave Technology*, vol. 38., no. 18, pp. 5116 - 5121, Sept. 15. 2020. DOI: 10.1109/JLT.2020.2997699
- [90] M. Tan *et al*, "Photonic RF arbitrary waveform generator based on a soliton crystal micro-comb source", *Journal of Lightwave Technology*, vol. 38, no. 22, pp. 6221-6226, Oct 22 (2020). DOI: 10.1109/JLT.2020.3009655.
- [91] M. Tan, X. Xu, J. Wu, R. Morandotti, A. Mitchell, and D. J. Moss, "RF and microwave photonic high bandwidth signal processing based on Kerr micro-comb sources", *Advances in Physics X*, Vol. 6, No. 1, 1838946 (2020).
- [92] X. Xu, *et al*., "Continuously tunable orthogonally polarized RF optical single sideband generator based on micro-ring resonators," *Journal of Optics*, vol. 20, no. 11, 115701. 2018.
- [93] X. Xu, *et al*., "Orthogonally polarized RF optical single sideband generation and dual-channel equalization based on an integrated microring resonator," *Journal of Lightwave Technology*, vol. 36, no. 20, pp. 4808-4818. 2018.
- [94] X. Wang, Y. Zhao, Y. Ding, S. Xiao, J. Dong, "Tunable optical delay line based on integrated grating-assisted contradirectional couplers". *Photonics Research* **6**, 880-886 (2018).
- [95] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278-2324 (1998).
- [96] O. L. Mangasarian, W. N. Street, W. H. Wolberg, Breast cancer diagnosis and prognosis via linear programming. *Operations Research* **43**, 570-577 (1995).
- [97] C. M. Bishop, *Neural networks for pattern recognition*. (Oxford university press, 1995).
- [98] <https://software.intel.com/en-us/articles/measuring-instruction-latency-and-throughput>
- [99] D.J. Moss *et al*., "Tunable dispersion and dispersion slope compensators for 10Gb/s using all-pass multicavity etalons", *IEEE Photonics Technology Letters*, vol. 15, no. 5, 730-732 (2003). DOI: 10.1109/LPT.2003.809921.
- [100] D.J. Moss *et al*., "Multichannel tunable dispersion compensation using all-pass multicavity etalons", paper TuT2 Optical Fiber Communications Conference, Anaheim (2002). Postconference Technical Digest (IEEE Cat. No.02CH37339). Opt Soc. America. Part vol.1, 2002, pp. 132-3. Washington, DC, USA.
- [101] D.J. Moss *et al*., "Tunable dispersion compensation at 10 Gb/s and 40 Gb/s using multicavity all-pass etalons", *Optical Fiber Communications Conference (OFC) paper TuD1*, page 162, Atlanta, GA, March (2003). Post-conference Digest (IEEE Cat. No.03CH37403). Opt. Soc. America. Part vol.1, 2003, pp. 162-3. Washington, DC, USA.
- [102] L. Lunardi *et al*., "Tunable dispersion compensators based on multi-cavity all-pass etalons for 40Gb/s systems", *Journal of Lightwave Technology*, vol. 20, no. 12, 2136 (2002). DOI: 10.1109/JLT.2002.806768.
- [103] L.M. Lunardi *et al*., "An Etalon-Based Tunable Dispersion Compensator Device for 40-Gbit/s Applications", *European Conf. on Optical Communications*, Paper 5.4.6 Copenhagen, Sept. (2002). vol. 2, 2002, Piscataway, NJ, USA. Print ISBN: 87-90974-
- [104] Mengxi Tan, Xingyuan Xu, Jiayang Wu, Thach G. Nguyen, Sai T. Chu, Brent E. Little, Roberto Morandotti, Arnan Mitchell, and David J. Moss, "Photonic Radio Frequency Channelizers based on Kerr Micro-combs and Integrated Micro-ring Resonators", *JOSarXiv.202010.0002*.
- [105] Mengxi Tan, Xingyuan Xu, Jiayang Wu, Thach G. Nguyen, Sai T. Chu, Brent E. Littl, Arnan Mitchell, Roberto Morandotti, and David J. Moss, "Orthogonally polarized RF optical single sideband generation with integrated ring resonators", *JOSarXiv.202010.0001*.
- [106] Mengxi Tan, Xingyuan Xu, Jiayang Wu, Thach G. Nguyen, Sai T. Chu, Brent E. Littl, Arnan Mitchell, Roberto Morandotti, and David J. Moss, "Orthogonally polarized RF optical single sideband generation with integrated ring resonators", *Journal of Semiconductors*, 2021, In Press.
<http://www.jos.ac.cn/article/shaid/4c314db6c6dcfe19c275efde2b99a22329ba658eda52233b8ab4b8a08b55a7bf>.