Supplementary Information for Silicon photonics enabled universal cross-scale tensor processing on chip

Hao Ouyang[1], Zilong Tao[2], Jie You[3, *], Yihao Luo[1], Hao Hao[4], Shiyin Du[2], Jun Zhang[2], Shengjie Tang[5], Haibin Lv[5], Xiaoping Liu[5], Xin Zheng[3], Xiangai Cheng[1], and Tian Jiang[4, *]

[*]Correspondence to: jieyou1991@hotmail.com

[*]Correspondence to: tjiang@nudt.edu.cn

[1]College of Advanced Interdisciplinary Studies, National University of Defense Technology, 410073 Changsha, P.R. China.

[2]State Key Laboratory of High Performance Computing College of Computer, National University of Defense Technology, 410073 Changsha, P.R. China.

[3]Defense Innovation Institute, Academy of Military Sciences PLA China, 100071 Beijing, P.R.China

[4]Institute for Quantum Science and Technology, College of Science , National University of Defense Technology, 410073 Changsha, P.R. China.

[5]School of Physical Science and Technology, ShanghaiTech University, 201210 Shanghai, China,

Supplementary Note 1: SiP-CSTP chip preparation and characterization

1.1 Fabrication process of the SiP-SCTP chip

Our SiP-CSTP chip was fabricated on a standard 200 mm SOI wafer at the Shanghai Institute of Microsystem and Information Technology. The processing platform is a 130 nm process technology node on a 220nm SOI platform compatible with COMS. The silicon layer has a thickness of 220 nm and the BOX layer of SiO_2 has a thickness of 3 μm and the cross-sectional schematic is shown in Fig. S1.1 a. The size of the fabricated chip is about 3 × 15 mm^2 and the working area of this research is around 1.5 × 4.5 mm^2 with a scale of 2×2 computing unit. Fig. S1.1 b,c show the optical images of our SiP-CSTP chip before and after packaging, respectively.

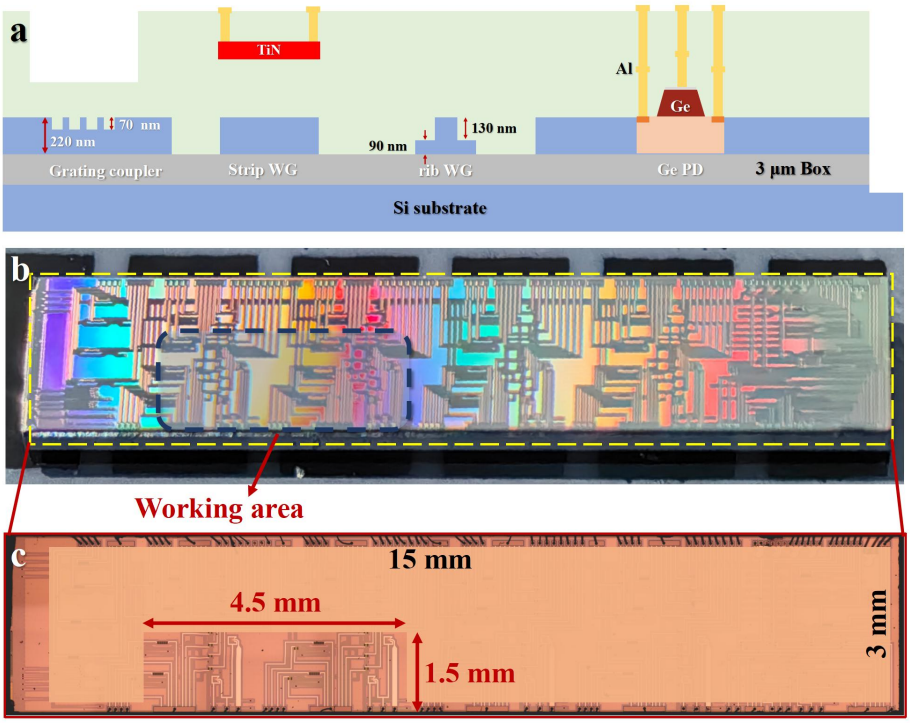

Figure S1.1. (a) The cross-sectional schematic of the wafer fab. The process includes on-chip key devices including gratings, waveguides, photodetectors and thermal phase shifters. (b) Optical images of unpacked SiP-CSTP chip with a footprint of 3 × 15 mm^2. (c) Microscopic synthesis images of packaged SiP-CSTP chip, where the employed computing area is only around 1.5 × 4.5 mm^2 in this work.

The SiP-CSTP chip comprises four primary components, namely the MRR for filtering functions, MZI for weight loading and computation, MRR group for WDM functionality, and BPD for receiving the computed optical signals. Additionally, the chip includes other auxiliary components such as

grating couplers for fiber-to-chip and chip-to-fiber laser coupling, beam splitters for power splitting, and optical signal monitoring ports. Figure S1.1 c illustrates the interconnection scheme among the devices. The signal light with loaded image information is transmitted through a filter and MZI in the computing group. Then, the output light from the upper and lower arms of MZI passes through WDM+ and WDM-, respectively and is input to the two optical ports of BPD to complete signal differentiation. The architecture, device connection sequence and the device number of the SiP-CSTP chip are shown in Figure S1.2. Researches has shown that, the performance of circuit is closely related to the performance of the device itself and the coordination between devices[1,2]. Generally, the system's performance would only reach its optimal level when the performance of each device reaches a good level and the parameter adaptation between each device is high.-

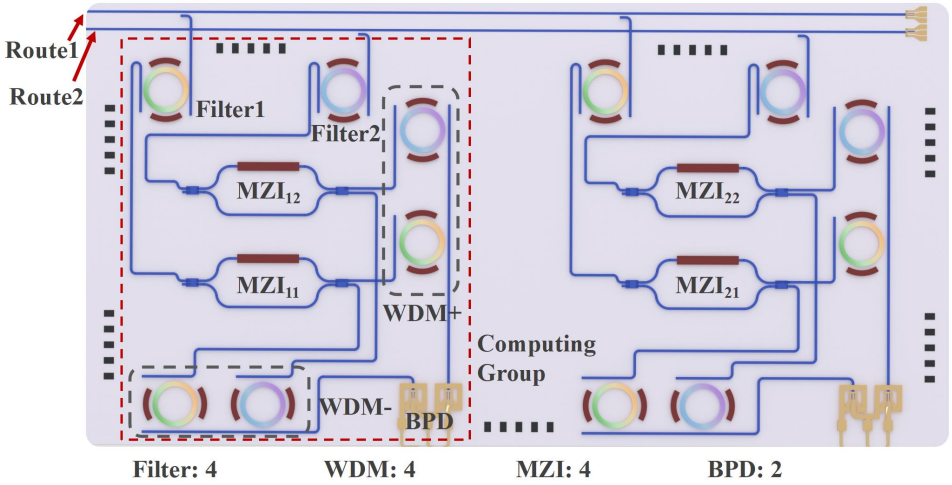

Figure S1.2. The architecture of SiP-CSTP chip, showcasing the various devices integrated within the chip, such as Filters, MZIs, WDMs, and BPDs..

1.2 Designed parameters of key devices on chip

The device performance greatly determines the system performance[3] which urges researchers to treat the system and architecture as a whole. One salient example are MRRs, which are one of the key components that are particularly sensitive to parameters in the CSTP system . Therefore, this section provides a detailed introduction to the design parameters of MRRs, while briefly introducing the design parameters of other devices.

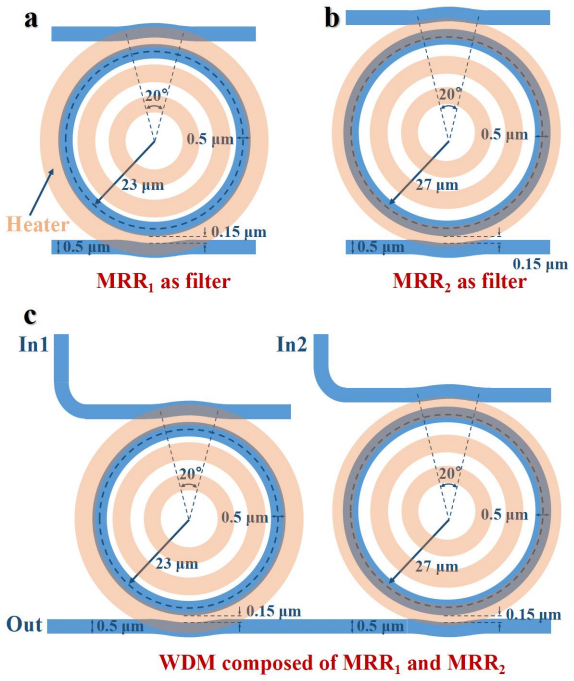

Figure S1.3. The designed parameters of filters and WDM based on MRRs. The diameters of MRRs are 23 and 27 μm respectively and the gap between bus and ring is 0.15 μm.

The diameters of the two types of MRRs are designed as 23μm and 27μm, corresponding to FSRs of ~8.0 nm and ~6.8 nm respectively for principal verification of the system scheme. The paths of the micro-rings are all based on the design of the 500 nm wide strip waveguides, and the basic mode near 1550 nm is TE mode. To achieve enhanced coupling efficiency, the spacing between the bus waveguide and the micro-ring is set at 150 nm. In order to make the resonant peak of the micro-ring adaptable to more working scenarios, it is necessary to design a TiN heating phase shifter to ensure that its resonant peak is actively adjustable. The path structure of the heating electrode includes multiple rings to improve heating efficiency. The parameter details of MRRs design are shown in Fig. S1.3 a and b. To ensure optimal wavelength selectivity and maintain consistency and completeness within our CSTP system's components, we have implemented an on-chip WDM structure, which is achieved through the cascade arrangement of MRRs, as illustrated in Fig. S1.3 (c). It is noteworthy that the introduction of multiple wavelengths occurs via the In1 and In2 ports of the WDM, where the combined optical signals of these wavelengths are multiplexed and output through the out port.

The simulation results of the MRR transmission spectrum are shown in Fig. S1.4. It is found from Fig. (a) that the wavelength of the resonant peak increases with the increase of power applied to the thermally phase shifter in the MRR with a diameter of 23 μm. The spectra of the Through and Drop

ports of MRR without applying voltage are shown in Fig. S1.4 (b). Similarly, the simulated results of MRR with 27 μm are given in Fig. S1.4 (c) and (d).

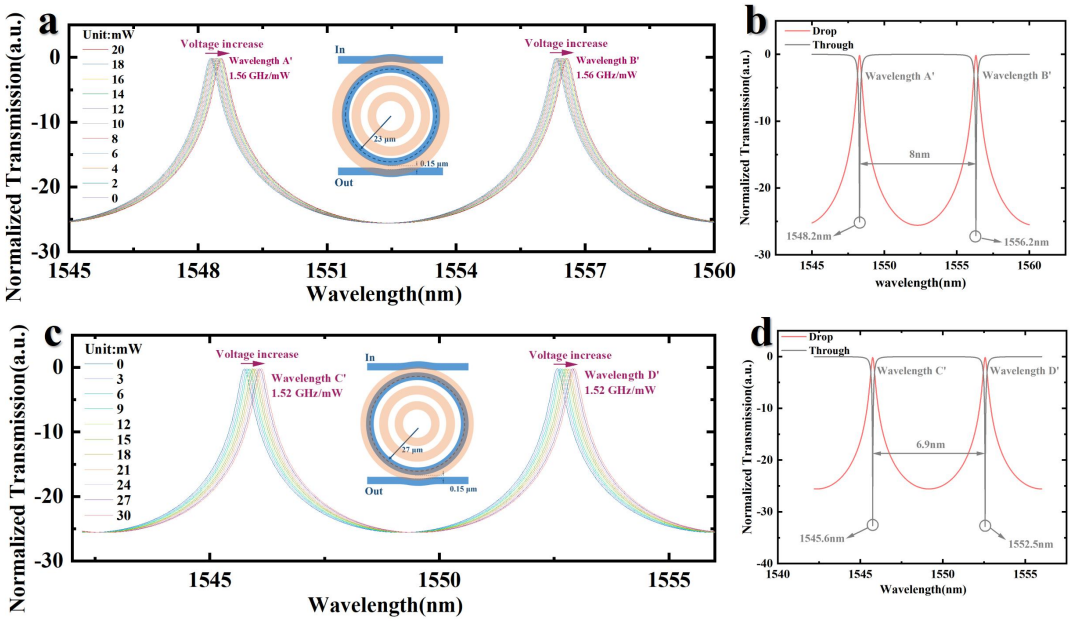

Figure S1.4. Simulation results for MRR with diameter of 23 and 27 μm. (a) Spectrum curves of MRR_1 (23 μm) with different voltages. (b) Spectrum curves of MRR_1 (23 μm) measured at Through and Drop port with non-voltage. (c) Spectrum curves of MRR_2 (27 μm) with different voltages. (d) Spectrum curves of MRR_2 (27 μm) measured at Through and Drop port with non-voltage.

Fig. S1.5 presents the characterization results of two MRR-related devices integrated in each of the two computing groups within the system. The normalized transmittance experimental results of the filter composed of MRR are shown in the four black curves in Fig. S1.5. Each resonant peak follows the Lorentz distribution, and the difference in transmittance peak intensity of all resonant peaks is within 3dB, indicating the stability of the filter for filtering in various wavelengths. The four resonant peak wavelengths required for our SiP-CSTP system are ~1548.3 nm (A'), ~1556.3 nm (B'), ~1545.8 nm (C') and ~1552.6 nm (D'), respectively. Additionally, the test results of the MRR-based WDM+ and WDM- are illustrated by the red and blue curves in Figure S1.5, respectively. In order to ensure proper operation of the WDM, it is crucial to prevent wavelength overlap between different routes within each computing group. Based on the test results, our WDM successfully meets this requirement, exhibiting good performance with an isolation degree of approximately 20 dB between different wavelengths.

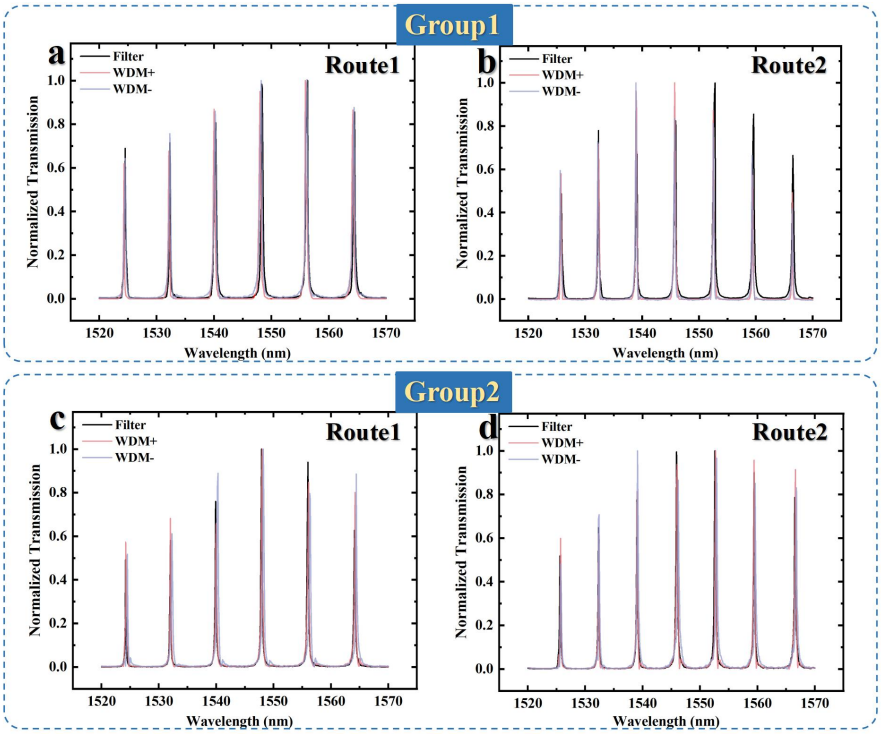

Figure S1.5. Transmission spectra of filters and WDMs consisting of MRRs at the drop ports. (a) Transmission curves of Filter (black), WDM+ (red), and WDM- (blue) in route 1 of computing group 1. (b) Transmission curves of Filter (black), WDM+ (red), and WDM- (blue) in route 2 of computing group 1. (c) Transmission curves of Filter (black), WDM+ (red), and WDM- (blue) in route 1 of computing group 2. (d) Transmission curves of Filter (black), WDM+ (red), and WDM- (blue) in route 2 of computing group 2.

Unlike traditional cascaded MZI architecture[4-6], CSTP system mainly utilizes the intensity modulation caused by the phase change of MZI to load the weight information, and needs to extend the loading number domain from the traditional positive number domain to the entire real number domain. So, in the MZI design process, BPD needs to be combined for joint design. Firstly, in terms of device parameters, as shown in Fig. 1.6. The monitor route is used to monitor output+ and output- at optical monitor port and the rate between the through and monitor is 9:1.

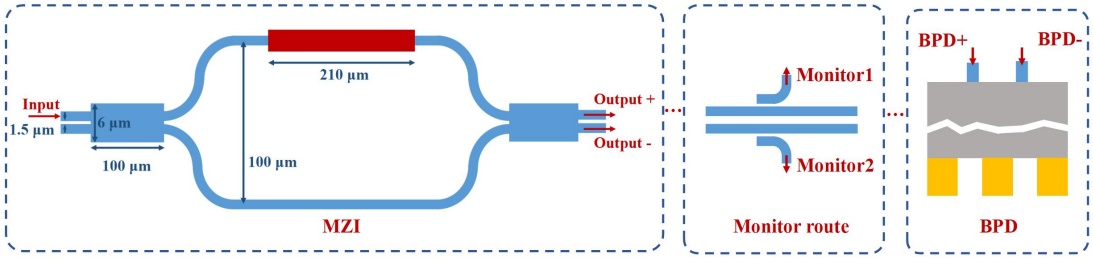

Figure S1.6. the diagram of MZI, Monitor, BPD integrated design.

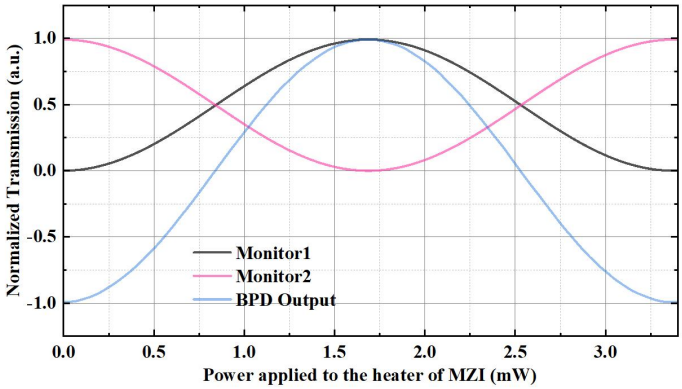

Figure S1.7. the simulated MZI normalized optical intensity variation curve with power applied to heater. The P_π of MZI is about 1.65 mW.

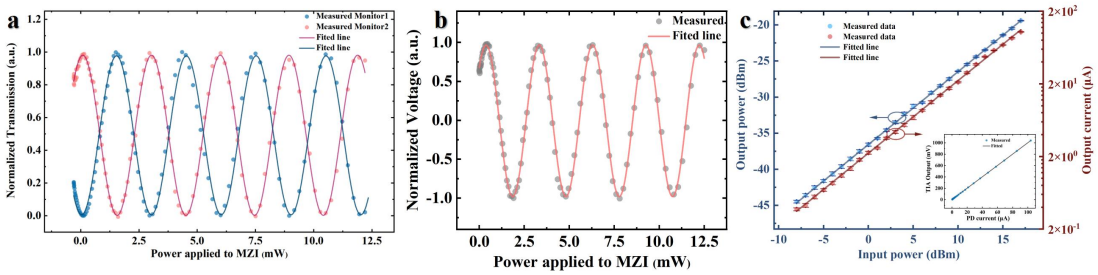

Figure S1.8. the measured results of MZI and BPD. (a) The measured normalized transmission changes with the power applied to heater of MZI. (b) Normalized differential output electrical signal in BPD after TIA. (c) The variation trend of BPD output power (left) and current (right) with input optical signal intensity in single channel.

The simulation results based on the MZI structure design are shown in Fig. S1.7, where the black and red curves represent the light intensity signals of the Monitor 1 and Monitor 2 ports after MZI output. The simulation results show that these two outputs exhibit strict sine and cosine signals, respectively, with a half wave power P_π of 1.65 mW. The differential results of the two optical signals output by the upper and lower arms of MZI were detected by BPD, and the results are shown in the blue curve. Due to the introduction of BPD, the amplitude of the output is expanded from only positive domain representation to the entire real number representation. Meanwhile, the amplitude intensity can also reach twice that of a single optical port in MZI.

The actual test results of the MZI in the CSTP chip are shown in Fig. S1.8 (a). It can be seen that the actual test results are almost consistent with the simulation results in terms of waveform and half wave power, but there is a significant difference in phase. This is mainly reflected in the imbalance between the upper and lower arms during the actual machining process and the different spectral ratios of the

MMI. To solve this, we will initialize and calibrate it in subsequent experiments. The output signals of the upper and lower channels of MZI are detected by BPD, as shown in Fig. S1.8 (b). The signal exhibits real number domain expression, which matches the simulation. The performance of photodetector is shown in Fig. S1.8 (c). As the input optical power of the bus changes, the blue curve represents the trend of the optical power at the optical monitoring port before the input of the BPD. Conversely, the red curve represents the corresponding changes in the output photocurrent of the BPD in response to the input optical power of the bus. The inset shows the relationship between the output voltage of the TIA and the photocurrent, which present the good linearity of PD.

Supplementary Note 2: SiP-CSTP system theoretical model and experimental results.

2.1 SiP-CSTP system theoretical model

Due to the fact that the CSTP system model contains a large number of core components, it is necessary to establish a model for each component including MRR, MZI and BPD. Based on the transmission order of light in the chip path, the MRR is first modeled. The MRR is in the form of a circular ring and two straight waveguides as a fraction of the input power coupled to the ring. The light will circulate through the ring and suffer a phase shift $\Delta\varphi = \beta L$, where β is the propagation constant in the ring, and L is the optical path length. The device will act as a resonator if this phase shift is an integral multiple of 2π. The resonance condition is given by[7]

$$\Delta\varphi = \beta L = 2m\pi \tag{2.1}$$

where m is an integer. Using $L = 2\pi R$, where R is the radius of the ring, and expressing β in terms of effective index *N*, the resonance condition may be written as

$$\lambda_m = \frac{2\pi NR}{m} \tag{2.2}$$

The analysis of the ring resonator may be performed by using coupled mode theory. The input and output electric fields of the straight guide are E_i and E_{th}, respectively. The electric field coupled into the ring is B_i and that after covering length *L* is B_0. And the electric fields of drop port in the ring is E_{dr}. Since we assume that the coupling coefficients of the two coupling regions are exactly the same, the steady-state input and output fields are written as

$$E_{th} = (1-\gamma)^{1/2}[E_i \cos(\kappa L_c) - jB_0 \sin(\kappa L_c)] \tag{2.3 a}$$

$$B_i = (1-\gamma)^{1/2}[-jE_i \sin(\kappa L_c) + B_0 \cos(\kappa L_c)] \tag{2.3 b}$$

$$E_{dr} = (1-\gamma)^{1/2}[-jB_i \sin(\kappa L_c)] \tag{2.3 c}$$

where κ is the mode-coupling coefficient, γ is the intensity insertion loss coefficient, and L_c is the coupling length. It is assumed that the propagation constants are the same, β, in straight and ring guides. If the intensity attenuation coefficient of the ring is denoted by α, B_0 is expressed as

$$B_0 = B_i \exp[(-\alpha L/2) - j\beta L] \tag{2.4}$$

The field transmittance of the ring resonator in through and drop port is then expressed by using Eqs. (2.3) and (2.4) as

$$\frac{E_{th}}{E_i} = (1-\gamma)^{1/2}\left[\frac{\cos(\kappa L_c) - (1-\gamma)^{1/2}\exp[(-\alpha L/2) - j\beta L]}{1-(1-\gamma)^{1/2}\cos(\kappa L_c)\exp[(-\alpha L/2) - j\beta L]}\right] \quad (2.5\text{ a})$$

$$\frac{E_{dr}}{E_i} = (1-\gamma)^{1/4}\left[\frac{\sin^2(\kappa L_c)\exp[(-\alpha L/4) - j\beta L/2]}{1-(1-\gamma)^{1/2}\cos(\kappa L_c)\exp[(-\alpha L/2) - j\beta L]}\right] \quad (2.5\text{ b})$$

Introducing new parameters defined as

$$x = (1-\gamma)^{1/2}\exp(-\alpha L/2)\ ,\ y = \cos(\kappa L_c)\ ,\ \phi = \beta L \quad (2.6)$$

The intensity transmittance of the optical ring resonator in the through-port signal T_{th} and drop-port signal T_{dr} are given by

$$T_{th}(\phi) = \left|\frac{E_0}{E_i}\right|^2 = (1-\gamma)\left[1 - \frac{(1-x)(1-y^2)}{(1-xy)^2 + 4xy\sin^2(\phi/2)}\right] \quad (2.7\text{ a})$$

$$T_{dr}(\phi) = \left|\frac{E_{dr}}{E_i}\right|^2 = \left[\frac{x\cdot(1-y^2)^2}{1+x^2y^2-2xy\cos(\phi)}\right] \quad (2.7\text{ b})$$

Herein, $\phi = \beta L$ and β are physical quantities highly dependent on wavelength, which is from Eqs. (2.2).

Because we are using the Drop port of the MRR, the maximum and minimum transmittances are given by the following expression:

$$T_{dr,\max} = \frac{x\cdot(1-y^2)^2}{(1+xy)^2} \quad (2.8\text{ a})$$

$$T_{dr,\min} = \frac{x\cdot(1-y^2)^2}{(1-xy)^2} \quad (2.8\text{ b})$$

It is noted from these equations that $x \approx y \approx 1$ should be satisfied in order to maximize $T_{\max}$ and to minimize $T_{\min}$. Near the resonant peak $\lambda_m = 2\pi NR/m$ at each wavelength, it changes in a Lorentz line curve. Therefore, we approximate the transmission spectrum curve of the MRR as the linear superposition state of the Lorentz function at each resonant peak, as follows:

$$T_{dr}(\lambda) = \sum_{m=1}^{\infty}\frac{1}{\pi}\cdot\frac{1/2\cdot\delta\lambda}{(\lambda-\lambda_m)^2 + (1/2\cdot\delta\lambda)^2} \quad (2.9)$$

Where, $\delta\lambda$ is connected to full width at half maximum (FWHM), which is given by

$$\delta\lambda = \frac{2(1+xy)^2}{\pi x(1-y^2)^2} \quad (2.10)$$

Therefore, the transfer function of any wavelength passing through the microring can be obtained according to Eqs. (2.9), which also provides a theoretical basis for us to input multiple wavelengths simultaneously in the same MRR.

Next, we will establish a physical model of the MZI structure in the system. Given the complexity of the MZI port relationship and the large number of input and output ports, the following will be derived using the transfer matrix method to make the derivation process more concise. For the example of the 2 × 2 MZI required for this system in Fig. S1.6, the transmission matrix between the output optical port and the input optical port is represented as follows

$$\begin{pmatrix} \boldsymbol{E}_{o+} \\ \boldsymbol{E}_{o-} \end{pmatrix} = \boldsymbol{M} \begin{pmatrix} \boldsymbol{E}_{i+} \\ \boldsymbol{E}_{i-} \end{pmatrix} \tag{2.11}$$

Among them, the $\boldsymbol{E}_{o+}$ and $\boldsymbol{E}_{o-}$ are electronic field intensity of MZI optical output+ and output- port, and $\boldsymbol{E}_{i+}$ and $\boldsymbol{E}_{i-}$ are electronic field intensity of MZI optical input+ and input- port. The transmission matrix M consists of three parts, two multimode interferometers (MMIs) and the phase shift introduced by the two arms of MZI, expressed as follows

$$\boldsymbol{M} = \begin{pmatrix} \cos(\kappa_2 z_2) & -\boldsymbol{i}\sin(\kappa_2 z_2) \\ -\boldsymbol{i}\sin(\kappa_2 z_2) & \cos(\kappa_2 z_2) \end{pmatrix} \begin{pmatrix} \exp[\boldsymbol{i}(\phi_0 + \Delta\phi)] & \boldsymbol{0} \\ \boldsymbol{0} & \exp(\boldsymbol{i}\phi_0) \end{pmatrix} \begin{pmatrix} \cos(\kappa_1 z_1) & -\boldsymbol{i}\sin(\kappa_1 z_1) \\ -\boldsymbol{i}\sin(\kappa_1 z_1) & \cos(\kappa_1 z_1) \end{pmatrix} \tag{2.12}$$

Where $\kappa_2 z_2$ and $\kappa_1 z_1$ are the coupling efficiency of MMI_2 and MMI_1 respectively, ϕ_0 is the inherent phase caused by the transmission process of the upper and lower arms of MZI. $\Delta\phi$ is the phase shift introduced by the heater, which is connected to the voltage applied to. According to the property of the MMI, $\kappa_1 z_1 = \kappa_2 z_2 = \pi / \boldsymbol{4}$. Besides, $\boldsymbol{E}_{i-} = \boldsymbol{0}$, then the normalized optical intensity of output+ and output- port in MZI is expressed by

$$\boldsymbol{I}_{o+}(\Delta\phi) = \sum_{s=1}^{s=m} \frac{\boldsymbol{I}_{i+}(\lambda_s)}{\boldsymbol{2}} [\boldsymbol{1} - \cos(\Delta\phi)] \tag{2.13 a}$$

$$\boldsymbol{I}_{o-}(\Delta\phi) = \sum_{s=1}^{s=m} \frac{\boldsymbol{I}_{i+}(\lambda_s)}{\boldsymbol{2}} [\boldsymbol{1} + \cos(\Delta\phi)] \tag{2.13 b}$$

Here, $\boldsymbol{I}_{o+}$ and $\boldsymbol{I}_{o-}$ are optical intensity of MZI optical output+ and output- port, $\boldsymbol{I}_{i+}$ is the optical intensity of MZI optical input+ port.

Due to the fact that the two output ports of MZI are connected to the two input ports of BPD and connected to a TIA, the output voltage result of BPD is as follows

$$V_{OUT} = R_T \cdot \frac{\eta e}{h\upsilon}(I_{O+} - I_{O-}) \tag{2.14}$$

Where, V_{OUT} is the output voltage of BPD after passing through TIA. η is the quantum efficiency of BPD, e is the amount of charge of the electron. h is the Planck constant, v is the frequency of corresponding photons, and the transimpedance gain of the BPD is denoted as R_T. By integrating Eqs. (2.9), (2.13), and (2.14), the modeling process of the system architecture involves deriving the comprehensive model of the CSTP system. This is achieved by considering the propagation of the modulated multi-wavelength signal light within the chip. Taking one of the computing groups as a typical example, the system can be described as follows:

$$V_{11}(\Delta\phi) = \sum_{S=1}^{m} R_T \frac{\eta e \lambda_S}{hc} \{ I_{mod}(\lambda_S) \cdot (1-\gamma_{GC}) \cdot (1-\gamma_{WG}) \cdot T_{dr,\ filter}(\lambda_S) \cdot T_{dr,\ WDM+}(\lambda_S) \frac{1}{2} [1-\cos(\Delta\phi)] -$$
$$I_{mod}(\lambda_S) \cdot (1-\gamma_{GC}) \cdot (1-\gamma_{WG}) \cdot T_{dr,\ filter}(\lambda_S) \cdot T_{dr,\ WDM-}(\lambda_S) \frac{1}{2} [1+\cos(\Delta\phi)] \} \tag{2.15}$$

It is worth noting that $V_{11}(\Delta\phi)$ only displays the output results of BPD1 in the first computing group for multiple wavelengths in route 1, while all output results of BPD1 need to be overlaid with the results in BPD1 for all routes. The final result can be expressed as $V_{1BPD}(\Delta\phi_1, \Delta\phi_2 \cdots\cdots) = \sum_n V_{1n}(\Delta\phi_n)$. And it is the same applies to BPD2.

2.2 Experimental results

Section 2.1 of SI provides the modeling situation of the CSTP system, while this section mainly provides the actual test results of the system. Fig. S2.1 shows the characterization results of the four core computing MZI components of the system, and conducts relevant stability tests. The figure shows the results of 50 consecutive test groups, with a total test duration of over 4 hours. From the figure, it can be seen that due to factors such as system jitter and thermal crosstalk, the test results during system operation accumulate certain errors over time, which is also the main source of limited system accuracy.

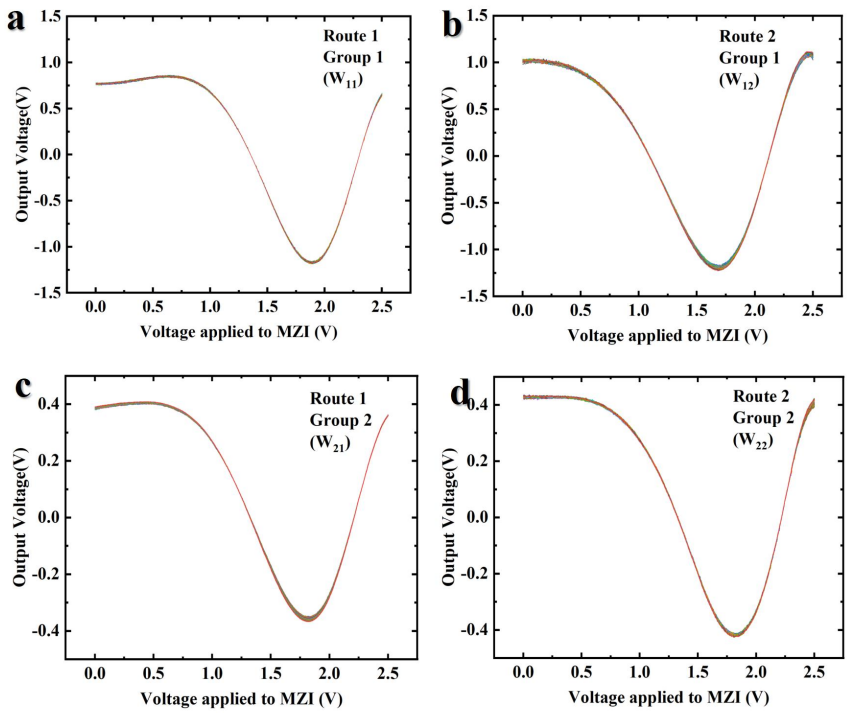

Figure S2.1. The relationship between the voltage loaded on MZI and the output voltage through BPD. The curve contains 50 sets of signal repeat test results. (a) the MZI signal presents W_{11} in Route1 Group1. (b) the MZI signal presented W_{12} in Route2 Group 1. (c) the MZI signal presents W_{21} in Route1 Group 2. (d) the MZI signal presented W_{22} in Route2 Group 2.

Based on similar evaluation schemes in the literature[8], the accuracy of the core calculation MZI components was evaluated, and we have introduced the definition. Weight control accuracy refers to the ability to achieve discrete levels in the weight database. For example, 10 bits accuracy means the ability to achieve 1024 distinguishable scatter levels. In our CSTP system weight library, assuming the measured voltage of weight is represented as V_i and the target weight is $\bar{V}_W$, the dynamic range can be expressed as the number of achievable discrete levels. The following will introduce the accuracy evaluation method of this system from a statistical perspective if V_i completed the following inequality:

$$\frac{1}{2^{n+1}} < \frac{\left|V_i - \bar{V}_W\right|}{\bar{V}_W} \leq \frac{1}{2^n} \tag{2.16}$$

Equation 2.16 represents the histogram of the test results for Vi, where the precision falls within n bits (n is a positive integer). Therefore, the number at n bits N_n in the histogram will increase by 1, where n and N_n are both positive. By applying Equation 2.16 iteratively, a statistical analysis of the 50,000 data points tested by the system yields the histogram of system precision, as shown in Fig. S2.2.

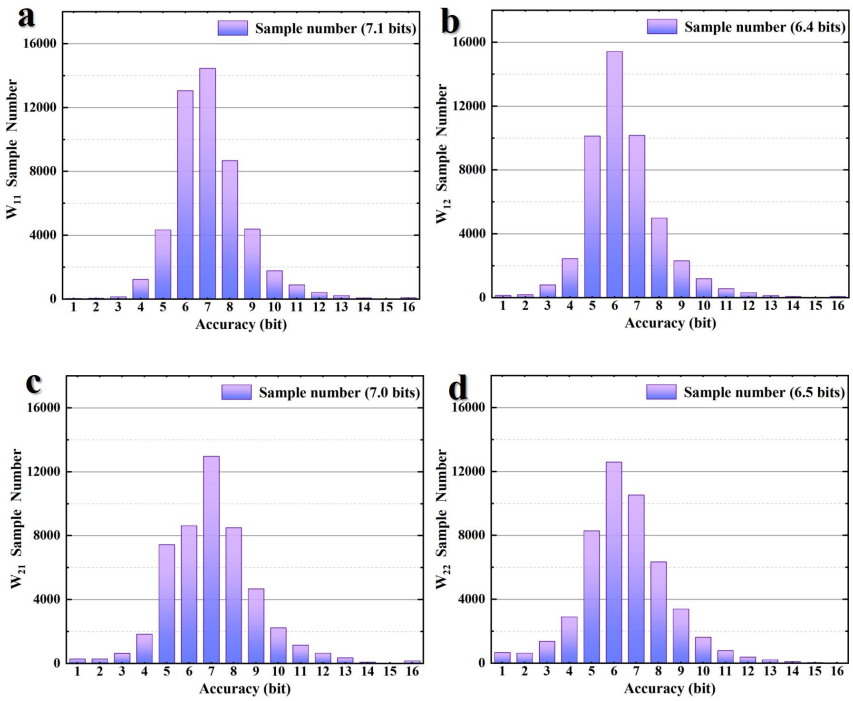

Figure S2.2. Accuracy statistical histograms of the four computing components MZI in the system. (a) the MZI signal presented W_{11} in Route1 Group1, 7.1 bits. (b) the MZI signal presented W_{12} in Route2 Group 1, 6.4 bits. (c) the MZI signal presented W_{21} in Route1 Group 2, 7.0 bits. (d) the MZI signal presented W_{22} in Route2 Group 2, 6.5 bits.

After obtaining the accuracy statistical histogram of each channel, the definition of the average accuracy of that channel is as follows

$$Accuracy = \frac{\sum_{n=1}^{\infty} N_n \cdot n}{\sum_{n=1}^{\infty} N_n} \tag{2.17}$$

Hence, the achieved accuracies for each channel are 7.1 bits (W_{11}), 6.4 bits (W_{12}), 7.0 bits (W_{21}), and 6.5 bits (W_{22}), respectively. Considering the presence of the barrel effect, the accuracy of all four channels in the system needs to be uniformly balanced. Consequently, the loading accuracy of the system's weight matrix is consistently set at 6 bits.

After completing the basic characterization, any tensor X and any weight matrix W are input into the CSTP system. The computed voltage signal, namely the computing results of the input vector X and matrix W, is collected at the output port of TIA. The results in the time domain are shown in Fig. 4 (k) and (I) respectively, representing simulation and experimental results with an accuracy of 95.7%.

Supplementary Note 3: Calibration process of the system

In order to quantify the relationship between data and voltage, we need to scan and calibrate the system. The general process is as follows:

Firstly, we quantify the calculation of a single channel. After cross group amplification, the collected calculated voltage is:

$$V = V_0 + A_{abc} \times F_{abc}(V_{xab}) \times S_{ad}(V_{wad}) \tag{3.1}$$

Among them, V_0 is the voltage generated by the dark current of the BPD. A_{abc} is the sum of the effects of the devices at the *c*-th wavelength of the *b*-th route in the *a*-th computing group on the light intensity. F_{abc} is the effect of the voltage applied to the modulator (corresponding one in the modulator array) at the *c*-th wavelength of the *b*-th route in the *a*-th computing group on the calculated voltage result, with the amplitude ranges from 0 to 1. S_{ad} is the effect of the voltage applied on the *d*-th MZI in the *a*-th computing group on the results, with an amplitude range from -1 to +1. The simulation results and preliminary experimental results have shown that V_0 has a minimal impact on other parameters in the design, so we can ignore it or adopt compensation treatment.

Then, we fix V_{xab} leading to $F_{abc}(V_{xab})$ as a constant and scan V_{wab}. By scanning the intensity of the input voltage and collecting data corresponding to the resulting voltage, a linear interval with good linearity is selected as the working interval of V_{wab}, and the results are normalized to obtain $S_{ad}(V_{wad})$. After collecting the curve of $S_{ad}(V_{wad})$, further set it to 1, and scan V_{xab} to obtain the data of the input voltage corresponding to the result voltage. Select a region with good linearity as the working interval of V_{xab}, and normalize the results to obtain $S_{ad}(V_{wad})$. At the same time, record the normalization factor to obtain the calculation result through the normalization factor.

For multi-channel situations, it is also necessary to maintain the same numerical value for the same computing area A_{abc}, which can be achieved by reasonably selecting the working intervals of V_{xab} and V_{wab}.

Supplementary Note 4: Convolution procedure and prediction results for Cifar-10 and Google Quickdraw recognition

To evaluate the efficacy of our model on different datasets, we conducted experiments on the MNIST dataset, the CIFAR-10 dataset and the Google QuickDraw dataset.

After 20 rounds of training, the convergence of our network on the MNIST dataset was assessed using cross-entropy as the loss function. Fig. S4.1 illustrates the progression of the training process. Remarkably, our model achieved near-convergence after the second epoch, indicating its rapid learning capability, as shown in Fig. S4.1 (a) and (b). Subsequent iterations primarily focused on fine-tuning the network parameters to further enhance its performance. As shown in Fig.4.1(c)-(d), the confusion matrix demonstrates the accuracy of prediction obtained from the experimental and theoretical calculation are 97.90% and 97.86%, respectively.

When it comes to the CIFAR-10 dataset, Fig. S4.2 (a) and (b) shows the change curves of the loss and accuracy during training process of the model. After 20 epochs of training, the model has converged sufficiently. The experimental and simulation results of confusion matrix, depicted in Fig. S4.2 (c) and (d), demonstrate the model's performance on the CIFAR-10 dataset with the accuracy of prediction obtained from the experimental of 97.90% and theoretical calculation of 97.86%. These figures illustrate a comprehensive analysis of the network's accuracy and validation loss, providing insights into its classification capabilities.

Furthermore, we also evaluated the performance of the model on the Google QuickDraw dataset by confusion matrix, as shown in Fig. S4.3. As shown in Fig. 4.3 (a) and (b), similar to the MNIST dataset, the model has essentially converged after the second epoch, and the subsequent iterations mainly involve fine-tuning. As depicted in Fig. 4.3 (c) and (d), the experimental and theoretical calculated prediction accuracies are 93.51% and 93.73%, respectively.

In a word, these figures present a detailed comparison between the experimental and simulated results, showcasing the model's ability to accurately recognize and classify hand-drawn sketches from various categories. By expanding the training process to encompass multiple datasets, our study provides a comprehensive assessment of the model's performance and its generalization capability across different data domains. These findings contribute to the advancement of silicon-based photonic computing research, showcasing the potential of our approach in achieving high accuracy and robustness in various applications.

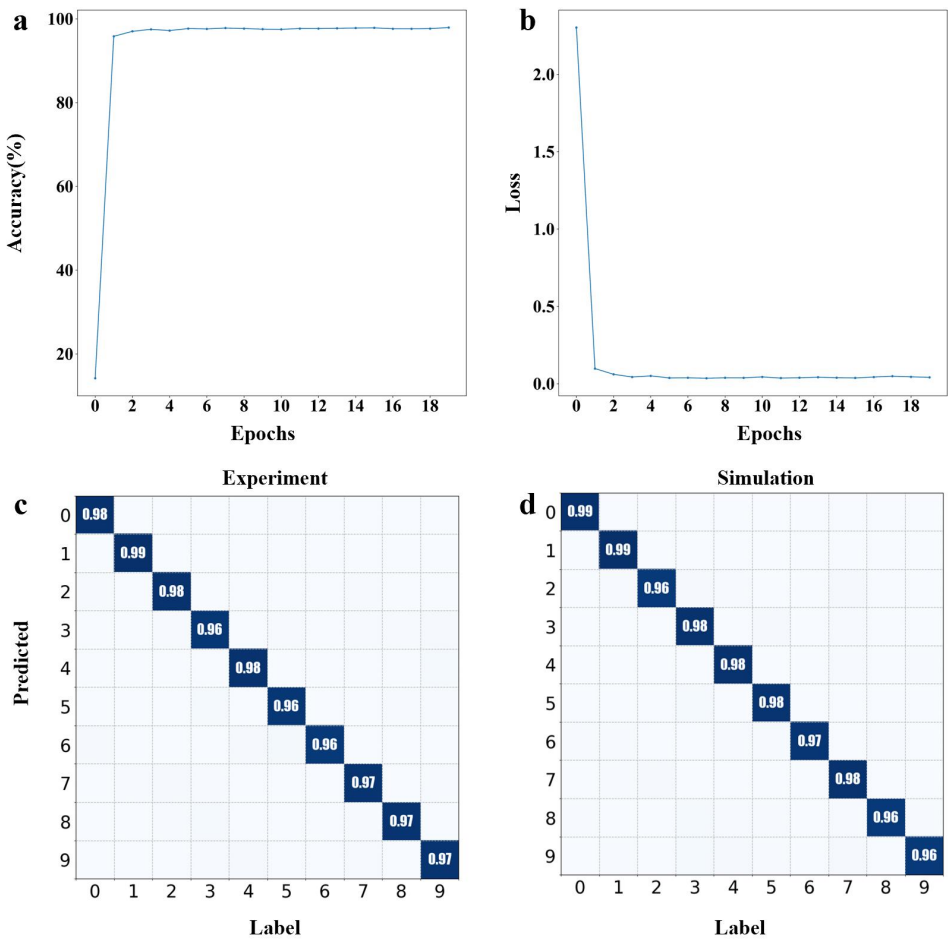

Figure S4.1. (a)-(b) The trend curves of accuracy and loss function during the training process on the MNIST dataset. (c)-(d) The confusion matrix of experiments and simulation calculations in the handwritten digit recognition task of the MNIST dataset, the total accuracies are 97.86% and 97.90%, respectively.

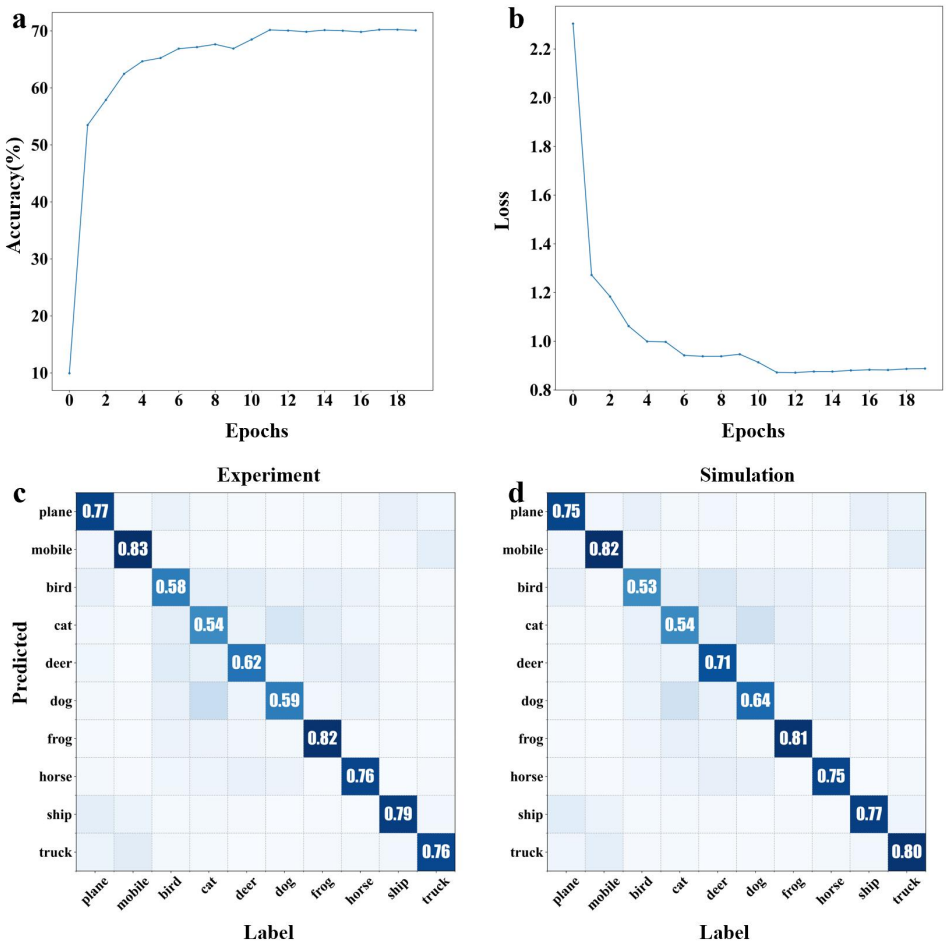

Figure S4.2. (a)-(b) The trend curves of accuracy and loss function during the training process on the CIFAR-10 dataset. (c)-(d) Confusion matrices for object recognition tasks on the CIFAR-10 dataset with experimental and simulated calculations, yielding overall accuracies of 70.22% and 70.70%, respectively.

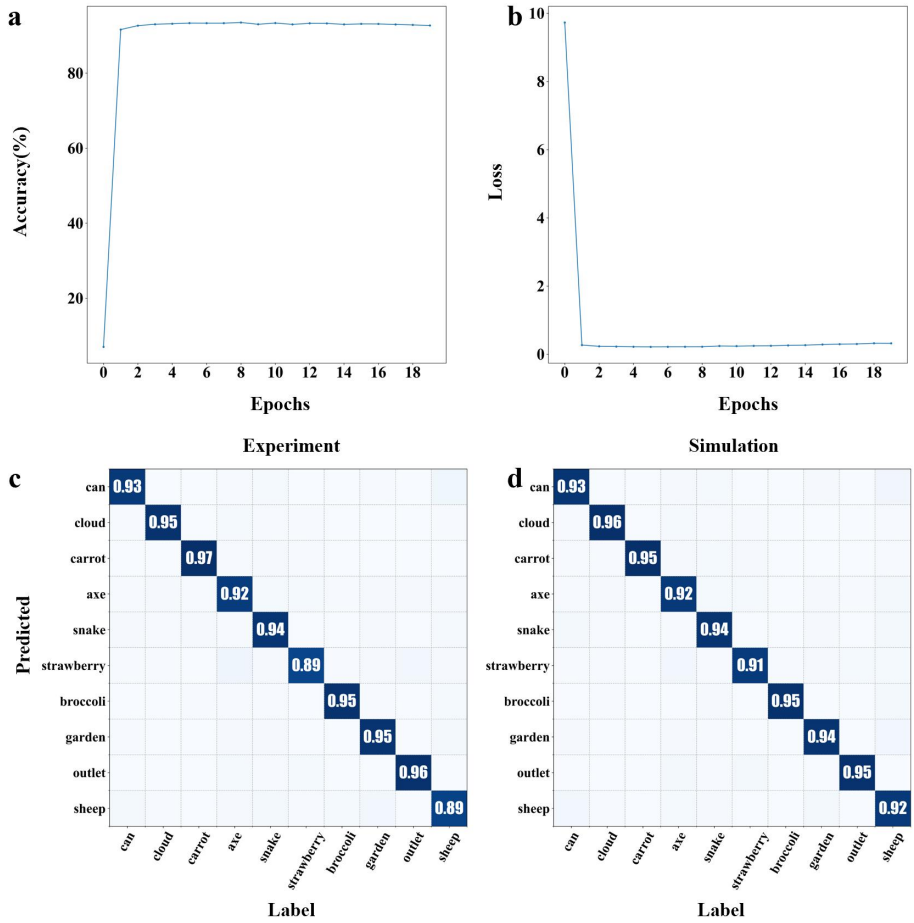

Figure S4.3. (a)-(b) The trend curves of accuracy and loss function during the training process on the Google Quickdraw dataset. (c)-(d) Confusion matrices for recognition tasks on the Google Quickdraw dataset, with experimental and simulated calculations, resulting in overall accuracies of 93.51% and 93.73%, respectively.

In this study, we employed a neural network model for training on three datasets: MNIST, CIFAR-10, and Google QuickDraw. The architecture of the neural network used for CIFAR-10 and Google QuickDraw datasets closely resembles that of the model used for the MNIST dataset. The main difference lies in the input layer, where the CIFAR-10 dataset comprises 3-channel input due to its RGB color images, while MNIST has single-channel grayscale images. Consequently, further elaboration on the network architecture's commonality is not necessary, as it remains consistent across the datasets.

The CIFAR-10 dataset, being a benchmark for image classification tasks, poses more complex challenges than the grayscale MNIST dataset. Using the identical neural network architecture, our

objective is to evaluate the model's ability to process multi-channel input and extract significant features from color images. This evaluation is of particular interest in the field of silicon-based photonic computing, where efficient processing of complex visual data is crucial for a wide range of applications, such as image recognition and object detection. Similarly, the Google QuickDraw dataset presents a unique set of challenges due to its diverse and hand-drawn sketch nature. By using the same neural network architecture as in the MNIST and CIFAR-10 datasets, we aim to demonstrate the model's versatility and adaptability to various types of data. This highlights the potential applicability of our approach in silicon-based photonic computing for diverse tasks, including image analysis, pattern recognition, and generative modeling. It is important to note that while the neural network architecture remains consistent, the training and optimization processes might have been tailored differently for each dataset to achieve optimal performance. These fine-tuning strategies might include adjusting learning rates, regularization techniques, and data augmentation methods, all of which contribute to improving the model's accuracy and generalization.

In conclusion, this study showcases the effectiveness and robustness of the neural network model across multiple datasets, demonstrating its potential for silicon-based photonic computing applications in handling diverse and complex data domains. The shared architecture allows us to efficiently adapt the model to different tasks, offering promising prospects for future advancements in the field.

Supplementary Note 5: Performance evaluation of SiP-CSTP system and parameters comparison of state-of-the-art integrated photonic computing hardwares.

5.1 Performance evaluation of SiP-CSTP system

To evaluate performance of our SiP-CSTP system, we define four figure-of-merits, including photonic *core compute density*, *overall compute density*, *energy efficiency estimation* and *computility per unit*. The following are descriptions of the above parameters.

Firstly, we define *photonic core compute density* as follows:

$$\textbf{\textit{Photonic core compute density}} = \frac{\textbf{\textit{Computing speed}}\ (\textbf{\textit{TOPS}})}{\textbf{\textit{Area of photonic linear operations unit}}\ (\textbf{\textit{mm}}^2)} \tag{5.1}$$

In this metric (also in line with the estimation protocols in ref [8-10]), the photonic core compute density of our SiP-CSTP chip is **0.9 TOPS mm^{-2}**, with computing power of 0.252 TOPS and area of photonic linear operations unit with four MZI units of $0.7 \times 0.1 \times 4 = 0.28$ mm^2.

Secondly, given the critical components involved, namely the photonic linear operations unit (MZIs), filters, WDMs, and BPDs, we employ all of them to conduct convolution operations. In this context, we define the *overall compute density* as follows:

$$\textbf{\textit{Overall compute density}} = \frac{\textbf{\textit{Computing speed}}\ (\textbf{\textit{TOPS}})}{\textbf{\textit{Total area of the photonic chips}}\ (\textbf{\textit{mm}}^2)} \tag{5.2}$$

Where, the detailed footprints of mentioned components above are listed in Tab. S5.1 to make a relatively comprehensive comparison, as well as the calculated overall compute density.

Table S5.1 Overall compute density estimation of the CSTP chip

Components	Footprint (mm^2)	number	Overall compute density (TOPS mm^{-2})
MZI	$0.7 \times 0.1 = 0.07$	4	$(4 \times 2 + 1) \times 2 \times 14 \times 10^9 / 1.644 \approx 0.153$
Filter	$0.045 \times 0.045 \approx 0.002$ [a]	4	
WDM	$0.3 \times 0.045 \approx 0.014$ [b]	4	
BPD	$0.65 \times 1.3 = 0.65$ [c]	2	
Total chip area	1.644		

[a] footprint of the filter is combined with a MRR.

[b] footprint of the WDM is combined with two MRRs with connecting waveguide.

[c] footprint of the BPD has great potential for reduction, it can be reduced to less than 0.3×0.4 mm^2, theoretically.

From the evaluation results, the total area in photonic chip is about 1.644 mm^2 and the corresponding overall compute density is about **0.153 TOPS mm^{-2}**. Strictly speaking, the area of the digital processing unit (CPU and GPU) and FPGA should be included in the compute density estimation. Unfortunately, the area of specific functional regions, particularly the computing core, lack publicly available information for reference. As a result, the estimation provided in this study does not consider the area of the digital processing unit and FPGA.

Thirdly, we define *energy efficiency* as follows:

$$\textbf{\textit{Energy efficiency}} = \frac{\textbf{\textit{Computing speed}}\ (\textbf{\textit{TOPS}})}{\textbf{\textit{Power}}\ (\textbf{\textit{W}})} \tag{5.3}$$

the power consumption of the current SiP-CSTP system are mainly from seven aspects: multi-wavelengths laser, EDFA, silicon photonic chip, modulator array, FPGA, CPU/GPU and TIA module as listed in Tab. S5.2

Table S5.2 Estimated power consumption of the CSTP system

Components	Voltage (V)	Current (mA)	Number	Power (mW)
Silicon photonic chip				~70.74
MZI	~ 1.5	~ 0.79	4	~ 4.74
Filter	~ 3.2	~ 1.68	4	~ 21.50
WDM	~ 3.2	~ 3.36	4	~ 43.00
BPD	1.25	~ 0.6	2	~ 1.5
Multi-wavelengths laser	/	/	1	~ 30, 000 [a]
EDFA	/	/	1	500 [a]
Modulator array	/	/	1	~ 6, 000 [a]
FPGA	/	/	1	~ 15,000 [b]
CPU / GPU	/	/	1	~ 25, 000 [c]
TIA module	5.0	~ 10	1	~ 50
Total power consumption				~ 76, 630.74

[a] typical power consumption according to the user manual.

[b] from a monitor software to obtain power consumption of FPGA.

[c] from a power monitor software when implementing fully-connected layers.

From the power consumption evaluation results, the total power consumption in photonic chip is about 76.63 W, corresponding to current energy efficiency of 3.28×10^{-3} TOPS/W. As a result of the dynamic changes in the calculated values on the SiP-CSTP chip during operation, it is necessary to

adjust the voltage applied to the thermal phase shifter of the MZI. This adjustment leads to varying power consumption of the chip. Multi-wavelengths laser, EDFA, modulator array, FPGA, CPU/GPU and trans-impedance amplifiers (TIA) module are the key devices off chip of the system, resulting their power consumption are relative higher than on chip device[9].

Thanks to the recent progress in photonic integrated circuits and devices, the numerous devices mentioned above have the opportunity to be implemented on the chip. Frequency combs (it can be integrated on chip and replace the multi-wavelengths laser)[11,12], circuit-based EDFA[13], on-chip semiconductor optical amplifiers (SOA)[8] and co-packaged TIA chips[9] can be fully integrated on several chips eventually. Moreover, as hybrid and monolithic integration technology advances, it becomes feasible to integrate the light source, silicon photonic circuit, and related electronic components, such as modulator drivers, TIAs, digital-to-analog converters (DACs), and analog-to-digital converters (ADCs), onto a single main-board or even within a single chip. This integration enables a more compact and efficient system design, facilitating seamless communication and interaction between different functional blocks. The convergence of these components on a single platform enhances the overall performance, reduces power consumption, and streamlines the manufacturing process.

To demonstrate the exceptional capabilities of our CSTP architecture, we provide the anticipated power consumption based on similar protocols as in [10,14]. The power required for generating the microcomb can be as low as 98 mW [12], while the on-chip SOA (replacing EDFA) typically consumes 390 mW [8]. The energy consumption associated with photodetection is primarily determined by the transimpedance amplifier (TIA). The power consumption from various digital circuits is as follows: a 5.36 pJ per sample driver with modulator (28 GHz) [15], a 1.14 pJ per sample TIA (53 GHz) [16], a 2.72 pJ per conversion DAC (8 bits) [17], and a 2 pJ per conversion ADC (8 bits) [18]. Considering the computing speed of the prototypical CSTP at 0.252 TOPS, the expected power consumption is calculated as $98 + 390 + (5.36 \times 4 + 1.14 \times 2 + 2.72 \times 2 + 2 \times 2) \times 14 = 952.24$ mW. Consequently, the anticipated energy efficiency of our prototypical CSTP system is approximately 0.265 TOPS/W. However, this is not the highest performance state of this system in the future. The above is only described from the perspective of reducing power consumption, there are several methods to improve energy efficiency, including increasing computing speed and expanding cross scale design. If the operating speed of the system remains at 14 Gbaud and the soft cross scale computing ability continues to be improved, as

shown in Tab. S5.3, then in the $W_{[4\times4]}\bullet X_{[4\times m]}$ scale case, the energy efficiency calculated according to the above equation is 1.97 TOPS/W. This showcases the remarkable efficiency and power-saving potential of our cross-scale architecture, highlighting its suitability for a wide range of applications.

Fourthly, in order to better demonstrate the effectiveness and ability of cross scale computing architecture in this work, we define *computility per unit* as follows:

$$\textbf{\textit{Computility per unit}} = \frac{\textbf{\textit{Computing speed}}\ (\textbf{\textit{TOPS}})}{\textbf{\textit{Computing unit}}} \tag{5.4}$$

The number of computing units is directly correlated to the number of MZIs integrated on our CSTP chip. The current SiP-CSTP chip consists of four MZIs as core computing unit presenting the weight matrix $W_{[2\times2]}$, which correspond to computility per unit of 0.063 TOPS/unit at 0.252 TOPS top computing power. The above unit computing power is estimated in the scale of 2 × 2, but the ultimate operational scale will reach to $W_{[i\times n]}\bullet X_{[n\times m]}$, where *m* represents the number of wavelengths in each route, *n* represents the number of routes in the chip, and *i* represents the number of computing groups. By increasing the number of optical wavelengths through the MRRs without increasing hardware costs, we can still further improve computing power and largely enhance computility per unit in the SiP-CSTP system. The computility per unit at different cross-scales computing are shown in Tab. S5.3.

Table S5.3 Estimated computing power and computility per unit of the CSTP system at modulator rate of 14 Gbaud.

Computing scale	*m*	*n*	*i*	Computing Power (TOPS)	Computing Power benchmark ratio [a]	Computility per unit (TOPS/unit)
$W_{[2\times2]}\bullet X_{[2\times m]}$	1	2	2	0.112 [b]	1	0.028
	2	2		0.252 [c]	2.25	0.063
	4	2		0.532	4.75	0.133
	8	2		1.092	9.75	0.273
$W_{[2\times4]}\bullet X_{[4\times m]}$	1	4	2	0.224 [d]	1	0.028
	2	4		0.476	2.125	0.060
	4	4		0.980	4.375	0.123
	8	4		1.988	8.875	0.249
$W_{[4\times2]}\bullet X_{[2\times m]}$	1	2	4	0.224 [e]	1	0.028
	2	2		0.504	2.25	0.063
	4	2		1. 064	4.75	0.133

	8	2		2.184	9.75	0.273
$W_{[4\times4]}\bullet X_{[4\times m]}$	1	4	4	0.448 [f]	1	0.028
	2	4		0.952	2.125	0.060
	4	4		1.960	4.375	0.123
	8	4		3.976	8.875	0.249
……				……	……	……

[a] the ratio of the computational power achieved through cross-scale operations to the baseline computational power without cross-scale operations.

[b] the baseline computing power of traditional architecture at $n = 2$, $i = 2$.

[c] the computing power by cross-scale operation of this work.

[d] the baseline computing power of traditional architecture at $n = 4$, $i = 2$.

[e] the baseline computing power of traditional architecture at $n = 2$, $i = 4$.

[f] the baseline computing power of traditional architecture at $n = 4$, $i = 4$.

The calculation power evaluation for cross scale operations in the above table is determined by the formula $[m\times(2n+1)-1]\times i$. From the data of the table, we can find three useful information. **Firstly**, when the computing scale and computing unit number n and i are determined, increasing the number of wavelengths in each route m will still improve computing power, which is the advantage of cross-scale operation. **Secondly**, the increase in computing power through expanding hardware scale is linear, and exponential growth cannot be achieved unless cross-scale operations are used. This conclusion can be drawn by comparing data sets $W_{[2\times2]}\bullet X_{[2\times1]}$, $W_{[2\times4]}\bullet X_{[4\times1]}$, $W_{[4\times4]}\bullet X_{[4\times1]}$ and $W_{[2\times2]}\bullet X_{[2\times1]}$, $W_{[2\times2]}\bullet X_{[2\times2]}$, $W_{[2\times2]}\bullet X_{[2\times4]}$, $W_{[2\times2]}\bullet X_{[2\times8]}$. **Thirdly**, when simultaneously increasing m, n, and i, the fastest increase in computing power is achieved by increasing m, followed by i, and the weakest increase is n. This conclusion can be gotten by comparing $W_{[2\times2]}\bullet X_{[2\times8]}$, $W_{[2\times4]}\bullet X_{[4\times4]}$, $W_{[4\times2]}\bullet X_{[2\times4]}$ and $W_{[4\times4]}\bullet X_{[4\times2]}$ (at the same mathematical computing scale, the computing power are 1.092, 0.980, 1.064, 0.952 TOPS, respectively). This also demonstrates the superiority of cross-scale operation.

5.2 Comparison of integrated photonic computing hardwares

SI 5.1 provides an overview of the current and future system performance based on the CSTP architecture and examines the potential areas for optimization in this system. This section primarily focuses on two types of comparisons: the comparison of performance parameters and the functional comparison between our proposed system and the existing advanced system. Our system benefits from

the high-speed signal loading and cross-scale operation design and has some advantages in computing power compared with previously reported MZI, MRR, and MZI + Delay works. Detailed parameter comparisons are provided in Tab. S5.4.

Table S5.4 Comparative analysis of key parameters in leading silicon photonics computing chip architectures and hybrid architectural solutions.

Technology	Data loading rate (Gbaud)	Weight precision (bits)	Energy efficiency (TOPS/W)	Operational scale (W×X)	Computility per Unit [a] (TOPS/Uni)	Accuracy of different Dataset			
						MNIST	Google QuickDraw	CIFAR-10	Others
MZI mesh[19]	—	8	—	$W_{[4\times4]}\times X_{[4\times1]}$	—	—	—	—	Vowel phonemes 76.7%
MZI mesh[20]	1×10^{-5}	4	—	$W_{[4\times4]}\times X_{[4\times1]}$	4.94×10^{-8}	90.5%	—	71%	—
MZI mesh [21]	25	1	0.48	$W_{[4\times4]}\times X_{[4\times1]}$	0.16	98.9%	—	—	—
Cascaded MZI [5]	1×10^{-7}	6.6	—	$W_{[1\times8]}\times X_{[8\times1]}$	1.88×10^{-10}	—	—	—	AUTO MAP
MRR[22]	0.047	> 5.5	—	$W_{[2\times2]}\times X_{[2\times1]}$	$\sim9.4\times10^{-5}$	—	—	—	—
MRR + WDM[3]	20	> 4.8	~ 0.06	$W_{[3\times4]}\times X_{[4\times1]}$	0.023	—	—	—	KTH
MRR + Deley[8]	17	9	0.2/ 1.52×10^{-3}	$W_{[1\times4]}\times X_{[4\times1]}$	0.034	96.6%	—	—	Edge detection
MRR + VOA[9]	—	—	~ 0.07	$W_{[1\times12}\times X_{[12\times12]}$;	$\sim 4.1\times10^{-3}$	—	—	—	Handwritten letter 93.8%
WDM + PCM[10]	2	5	~ 0.4	$W_{[4\times4]}\times X_{[4\times4]}$	~ 0.023	95.3%	—	—	Edge detection
Diffrative cell + MZI[23]	$\sim1\times10^{-5}$	—	~ 0.11	$W_{[10\times10}\times X_{[10\times1]}$	1.93×10^{-3}	80.4%	—	—	Iris flower 97.3%
Diffrative cell[24]	—	—	1×10^{-3}	—	—	98%	93%	45%	—
This work	14	6	0.27 / 1.97/ 3.28×10^{-3} [b]	$W_{[2\times2]}\times X_{[2\times2]}$ [d]	0.250 / 0.06 [c]	97.9%	93.5%	70.2%	N/A

[a] define Computility per Unit as the ratio of the total computing power to the number of computing units in the computing system. Detailed definition is exhibited in S5.1.

[b] refers to expected/current energy efficiency. 0.27 is obtained at the scale of $\boldsymbol{W}_{[2\times2]}\bullet\boldsymbol{X}_{[2\times2]}$, 1.97 is obtained at the scale of $\boldsymbol{W}_{[4\times4]}\bullet\boldsymbol{X}_{[4\times8]}$.

[c] refers to expected/current computility per unit. 0.250 is obtained at the scale of $\boldsymbol{W}_{[4\times4]} \cdot \boldsymbol{X}_{[4\times8]}$

[d] refers to cross-scale operation.

From the table, thanks to the CSTP architectural scheme, this work can expand the dimensions of tensor processing without increasing hardware, which is different from the method of other architecture to increase hardware scale. The computing power of our system increases nonlinearly with the number of wavelengths *m* in each channel, and the other part is the MZI computation group loading the weight matrix, which need not to change under cross-scale operation. The next step is to introduce optical frequency comb sources in the wavelength range of 1520 nm to 1570 nm, which can currently achieve a single-channel input of more than 8 wavelengths. If the feedback control of the micro-ring and light source parameters is more precise, allowing for a greater number of light sources to be simultaneously input into the micro-ring, theoretically, dozens of channels of the optical frequency comb can be achieved. Meanwhile, the difficulty of control also needs to be considered at the same time. Thus, this SiP-CSTP system allows for parallel processing of multiple matrix and vector operations, enabling low-cost acceleration for average pooling and channel fusion in convolutional neural networks, which are two key core functionalities of the SiP-CSTP chip. As shown in Tab. S5.5, compared to the computational functionality achieved by previous MZI, MRR, and diffractive architectures, our system offers a more comprehensive implementation of most processes in CNN, leveraging the advantages of silicon photonic chips. By utilizing the SiP-CSTP chip's acceleration capabilities in these two convolution processes, the classification recognition tasks of three datasets, MNIST, Google QuickDraw, and CIFAR-10, have been achieved experimentally. And the results are quite impressive compared to the recognition accuracy reported in previous literature, with accuracies reaching 97.5%, 93.5%, and 70.2% respectively.

Table S5.5 Comparison of functions and coding methods of different silicon photonics technical solutions.

Technology	Function				Encoding mode
	Pooling	Channel fusion	Convolution /MVM	Non-linear activation	
MZI mesh [20]	—	—	●	ReLU (Electrically post-processed)	Implicit coding
MRR[22]	—	—	●	ReLU (On chip)	Explicit coding
Diffrative cell + MZI[23]	—	—	●	Nonlinear activation function *f* (On chip)	Implicit coding
Diffrative cell[24]	●	—	●	ReLU (Electrically post-processed)	diffraction
This work	●	●	●	ReLU (Electrically post-processed)	Explicit coding

Therefore, based on the comparison of key parameters and relevant conclusions in the above table, it is demonstrated that the proposal of a CSTP architecture plays an important role in improving the four core indicators (*core compute density*, *overall compute density*, *energy efficiency estimation* and *computility per unit*) of the system. The upgrade and optimization of the architecture is very important in the next step of computing power improvement process. The three dimensions can effectively improve computing power, and are sorted as the number of wavelengths in each route *m*, the number of routes in the chip *n*, and the number of computing groups *i*. It is worth noting that the increase in *m* is bound to the results of cross scale operations and is also the key to improving computility per unit. In addition to enhancing computing power, another approach to increase energy efficiency lies in reducing power consumption. In future developments, this system holds the potential to effectively address this by implementing on-chip designs that integrate multiple devices, resulting in decreased power consumption and improved overall energy efficiency.

References

1 Shu, H. *et al.* Microcomb-driven silicon photonic systems. *Nature* **605**, 457-463, (2022).

2 Li, J. *et al.* Fully integrated hybrid microwave photonic receiver. *Photonics Res* **10**, (2022).

3 Xu, S., Wang, J., Yi, S. & Zou, W. High-order tensor flow processing using integrated photonic circuits. *Nat Commun* **13**, 7970, (2022).

4 Prabhu, M. *et al.* Accelerating recurrent Ising machines in photonic integrated circuits. *Optica* **7**, (2020).

5 Xu, S. *et al.* Optical coherent dot-product chip for sophisticated deep learning regression. *Light Sci Appl* **10**, 221, (2021).

6 Mennea, P. L. *et al.* Modular linear optical circuits. *Optica* **5**, (2018).

7 M. Jamal Deen, P. K. B. in *Silicon Photonics* 303-331 (2012).

8 Bai, B. *et al.* Microcomb-based integrated photonic processing unit. *Nat Commun* **14**, 66, (2023).

9 Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501-506, (2022).

10 Feldmann, J. *et al.* Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52-58, (2021).

11 Chang, L., Liu, S. & Bowers, J. E. Integrated optical frequency comb technologies. *Nat Photonics* **16**, 95-108, (2022).

12 Stern, B., Ji, X., Okawachi, Y., Gaeta, A. L. & Lipson, M. Battery-operated integrated frequency comb generator. *Nature* **562**, 401-405, (2018).

13 Liu, Y. *et al.* A photonic integrated circuit-based erbium-doped amplifier. *Science* **376**, 1309-1313, (2022).

14 Xu, X. *et al.* 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44-51, (2021).

15 Temporiti, E. *et al.* in *2016 IEEE International Solid-State Circuits Conference (ISSCC).* 404-405 (IEEE).

16 Lakshmikumar, K. R. *et al.* A process and temperature insensitive CMOS linear TIA for 100 gb/s/$\lambda $ PAM-4 optical links. *IEEE Journal of Solid-State Circuits* **54**, 3180-3190, (2019).

17 Kossel, M. A. *et al.* in *2021 IEEE International Solid-State Circuits Conference (ISSCC).* 130-132 (IEEE).

18 Kull, L. *et al.* A 24–72-GS/s 8-b time-interleaved SAR ADC with 2.0–3.3-pJ/conversion and> 30 dB SNDR at Nyquist in 14-nm CMOS FinFET. *IEEE Journal of Solid-State Circuits* **53**, 3508-3516, (2018).

19 Shen, Y. *et al.* Deep learning with coherent nanophotonic circuits. *Nat Photonics* **11**, 441-446, (2017).

20 Zhang, H. *et al.* An optical neural chip for implementing complex-valued neural network. *Nat Commun* **12**, 457, (2021).

21 Tian, Y. *et al.* Scalable and compact photonic neural chip with low learning-capability-loss. *Nanophotonics* **11**, (2021).

22 Huang, C. *et al.* A silicon photonic–electronic neural network for fibre nonlinearity compensation. *Nature Electronics* **4**, 837-844, (2021).

23 Zhu, H. H. *et al.* Space-efficient optical computing with an integrated chip diffractive neural network. *Nat Commun* **13**, 1044, (2022).

24 Hu, Z. *et al.* High‐Throughput Multichannel Parallelized Diffraction Convolutional Neural Network Accelerator. *Laser Photonics Rev* **16**, (2022).