RESEARCH ARTICLE Open Access

Monolithically integrated asynchronous optical recurrent accelerator

Bo Wu[1†], Haojun Zhou[1†], Junwei Cheng[1], Wenkai Zhang[1], Shiji Zhang[1], Chaoran Huang[2], Dongmei Huang[3], Hailong Zhou[1*], Jianji Dong[1*] and Xinliang Zhang[1,4*]

Abstract

Computing with light is widely recognized as a promising paradigm for overcoming the energy and latency limitations of electronic computing. However, the energy consumption and latency in current optical computing hardware predominantly arise in the electrical domain rather than the optical domain, primarily due to frequent signal conversions between optical (analog) and electrical (digital) formats. Furthermore, as the operating frequency of optical computing surpasses the GHz range, the synchronization of parallel electrical signals and the management of optical delays become increasingly critical. These challenges exacerbate energy consumption and latency, particularly in recurrent optical operations. To address these limitations, we propose a novel asynchronous computing paradigm for on-chip optical recurrent accelerators based on wavelength encoding, effectively mitigating synchronization challenges. By leveraging the intrinsic causality of wavelength relay, our approach eliminates the need for rigorous temporal alignment. To demonstrate the flexibility and efficacy of this asynchronous paradigm, we present two advanced recurrent models—an optical hidden Markov model and an optical recurrent neural network—monolithically integrated for the first time. These models incorporate hundreds of linear and nonlinear computing units densely packed into a compact footprint of just 10 mm^2. Experimental evaluations on various benchmark tasks underscore the superior energy efficiency and low latency of the proposed asynchronous optical accelerators. This innovation enables the efficient processing of large-scale parallel signals and positions optical processors as a pivotal technology for applications such as autonomous driving and intelligent robotics.

Keywords Asynchronous operation, Optical recurrent accelerator, Optical hidden Markov model, Optical recurrent neural network

1 Introduction

The rapid advancements in large language models and artificial intelligence present significant challenges to current electronic computing architectures, particularly in terms of energy consumption and latency. Due to the low propagation loss of photons in optical waveguides and the broad bandwidth of optical devices, optical computing has emerged as a promising alternative to alleviate these bottlenecks in electronic computing [1, 2]. However, it is widely acknowledged that optical computing also faces its own energy consumption challenges [3]. Specifically, the analog nature of optical computing necessitates the use of electrical digital-to-analog (DAC) and analog-to-digital (ADC) converters for data

†Bo Wu and Haojun Zhou have contributed equally to this work.

*Correspondence:
Hailong Zhou
hailongzhou@hust.edu.cn
Jianji Dong
jjdong@hust.edu.cn
Xinliang Zhang
xlzhang@mail.hust.edu.cn
[1] Wuhan National Laboratory for Optoelectronics, School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074, China
[2] Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
[3] Photonics Research Institute, Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Shatin, Hong Kong SAR, China
[4] Xidian University, Xi'an 710071, China

loading and retrieval, which dominate the energy consumption in typical optical computing systems [4]. To address this issue, substantial efforts have been directed toward designing architectures and applications that minimize reliance on conversions between analog and digital domains. One prominent solution is near-sensor computing, which places optical computing hardware closer to sensors, thus reducing the energy overhead associated with digital transitions [4, 5]. Additionally, the power consumption of DACs and ADCs is directly tied to their bit precision and working bit rate. For instance, as demonstrated in blockchain and cryptocurrency applications, low-precision sampling can reduce ADC power consumption and enhance robustness to errors [6].

Optical recurrent accelerators, which repeatedly utilize fixed optical operators, offer higher computing density compared to multilayer forward optical networks. Existing optical recurrent accelerators include applications such as optical reservoir computing [7, 8], Ising machines [9–11], and matrix inversion solvers [12, 13]. To reduce energy consumption and latency, it is desirable for optical recurrent accelerators to operate without intermediate electrical relays involving ADCs and DACs. However, many existing approaches, such as the Ising machine, often rely on electrical feedback to complete large-scale iterations, while optical reservoir computing is typically limited to single or cascaded fiber loops, lacking advanced parallel signal control [14]. The optical computing generally handles fast-varying input signals alongside slow-varying system parameters. A critical limitation of these systems is the synchronization challenge in managing multiple high-speed parallel input optical signals. While often overlooked during the construction of photonic computing cores, this issue becomes prominent as working frequencies exceed tens of gigahertz—necessary to achieve computing power surpassing current electronic computing cores [15]. Synchronization in this context encompasses the alignment of parallel input digital-to-analog (DA) signals, delay management across multiple optical links, and parallel ADC sampling of outputs. In electrical systems, synchronization relies on intricate control mechanisms, which add to energy consumption and latency [16–18]. In optical systems, synchronization requires maintaining identical delays across multiple optical paths in large-scale networks, greatly increasing the complexity of the optical network layout. This issue becomes particularly critical in optical recurrent accelerators, where even minor misalignments in optical signals can accumulate during iterative processes, potentially resulting in catastrophic disruptions to time-sequenced outputs (see Supplementary Information S1). Additionally, DACs must operate at precisely tuned data frequencies to match optical delays, necessitating higher sampling rates and further complicating synchronization efforts. The reliance on high-speed transistor-based circuits (TBCs) such as ADCs, DACs, drivers, and transimpedance amplifiers (TIAs) compounds these challenges because the power consumption of TBCs scales with their sampling frequency and can increase quadratically in optical transmission links [19–21]. Moreover, high-speed radio frequency (RF) signals experience significant losses in parasitic circuits, while their bit resolution tends to degrade at higher frequencies, adversely affecting computing accuracy [22]. These limitations highlight the urgency of developing a synchronization-free computational paradigm for advanced optical recurrent accelerators.

To address this challenge, we propose a unique asynchronous computation paradigm for on-chip optical recurrent accelerators. By mapping time sequences to optical wavelength sequences and leveraging an efficient on-chip wavelength relay, our approach eliminates the need for synchronization among input signals, optical paths, and output signals. This paradigm allows high-speed TBCs to be replaced by low-speed counterparts without increasing computing latency, thereby reducing the energy consumption and cost associated with RF signal generation, transmission, and reception. To demonstrate the versatility of our approach, we have designed and fabricated two monolithically integrated optical recurrent accelerators: an optical hidden Markov model (HMM) and an optical recurrent neural network (RNN). These accelerators incorporate hundreds of passive and active optoelectronic devices within a compact footprint of 10 mm^2. Performance evaluations on benchmark tasks highlight the exceptional energy efficiency and low latency of the asynchronous optical recurrent architecture, making it a promising solution for large-scale parallel signal processing in advanced applications.

2 Results

2.1 Principle of the asynchronous optical recurrent accelerator

The general representation of a recurrent system can be written as an iterative equation:

$$\vec{y}_{n+1} = F\left(\vec{x}_n, \vec{x}_{n-1} \cdots, \vec{x}_{n-k+1}\right), \tag{1}$$

where $\vec{x}_n$ represents the input vector at time n, $\vec{y}_n$ is the resulting vector at time n, and F denotes a fixed function. The output at time n is determined by the input vectors over the last k iterations, with the dimensionality of the vectors indicating the number of parallel channels involved in the iteration. The function F can either be a simple linear matrix multiplication or a more complex nonlinear operator. Traditionally, linear operations, such

as matrix–vector multiplication, are performed in the optical domain, while nonlinear operations are completed electrically [23]. Optical computing cores (OCCs) encompass diverse architectures, including those based on interference (e.g., Mach–Zehnder interferometer (MZI) meshes), diffraction (e.g., on-chip diffractive networks), and nonlinearity [24, 25]. These architectures typically serve as coprocessors for electronic hardware, where synchronization is managed. However, frequent interactions between optical and electronic domains reduce the potential benefits of optical hardware, such as latency and energy efficiency. To fully leverage the advantages of optical computing, all operations in the iterative equation can be meticulously mapped to photonic hardware. Figure 1a illustrates a typical high-speed synchronous architecture where DACs generate input signals that are amplified by drivers, converted to optical signals via modulators, and detected as photocurrents and converted to voltages by

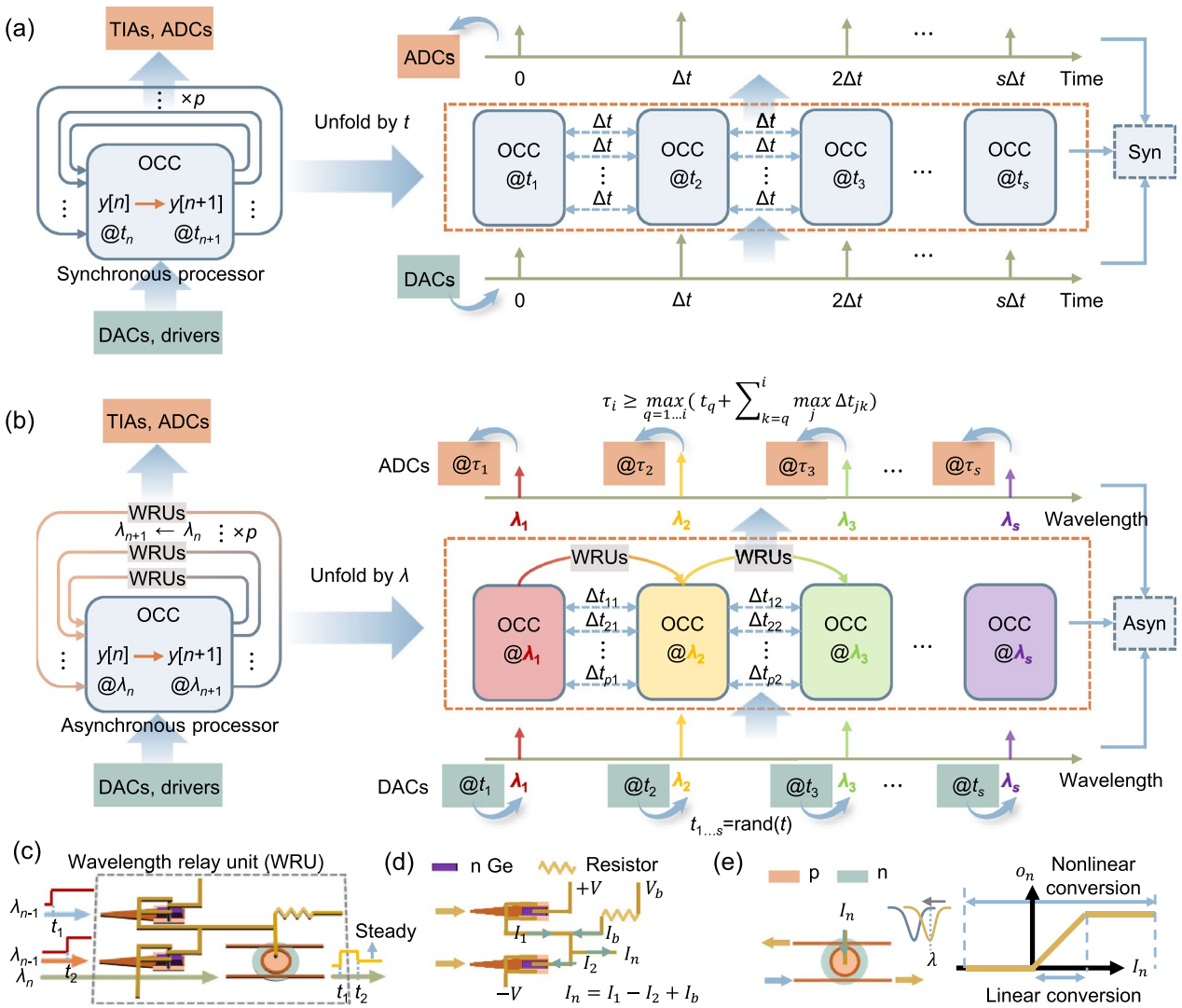

Fig. 1 Principle of the proposed asynchronous optical recurrent accelerator. **a** System diagrams of the optical synchronous processor. The computing process is unfolded cycle by cycle in the time domain, with the OCC operating in a time-multiplexed mode. ADCs and DACs operate at high speed, varying according to the period Δt. OCC, optical computing core. **b** System diagrams of the optical asynchronous processor. The computing process is unfolded cycle by cycle in the wavelength domain, with the OCC operating in a wavelength-multiplexed mode. ADCs and DACs operate in a quasi-static mode, generating or sampling a single electrical level asynchronously at $t_{1\ldots s}$ and $\tau_{1\ldots s}$. WRU, wavelength relay unit. **c** Information relay based on optical-electrical-optical conversion. The asynchronous input wavelength λ_{n-1} is differentially detected by photodetectors, which subsequently drive the MRM with the supply light of λ_n. The output signal stabilizes only after all signals from preceding cycles have arrived. **d** Electric current flow in the WRU. **e** Transfer function of the WRU, showing selectable linear and nonlinear working regions to accommodate specific requirements

trans-impedance amplifiers (TIAs). These signals are then sampled by high-speed ADCs. For synchronous operation, p parallel channels are required to maintain an identical delay Δt, which matches the data frequency of the DACs. However, due to the inherent challenges of signal synchronization, advances in parallel synchronous optical recurrent accelerators remain limited.

Our proposed asynchronous optical recurrent accelerator is illustrated in Fig. 1b. In this architecture, s cycles are mapped onto s distinct wavelength channels, and information transferred sequentially across cycles through a wavelength relay mechanism. To clarify the differences between synchronous and asynchronous computing architectures, we analyze the computing process cycle by cycle—using the time domain for synchronous processors and the wavelength domain for asynchronous processors. In the synchronous architecture, time-multiplexed OCCs are used, requiring high-speed DACs and ADCs that operate synchronously with the data frequency. In contrast, the OCC in the asynchronous architecture is broadband and wavelength-multiplexed. Each DAC and ADC operates in a quasi-static mode, generating or sampling a single electrical level rather than continuously handling high-frequency signals. This asynchronous operation relies on the inherent information causality within the wavelength sequence, enabling asynchronous loading of signals and asynchronous detection of results across different wavelength channels. Consequently, the strict temporal alignment required in synchronous systems is eliminated, allowing each optical path to have a variable length. The detection time only needs to satisfy the condition:

$$\tau_i \geq \max_{q=1\ldots i}\left(t_q + \sum\nolimits_{k=q}^{i} \max_j \Delta t_{jk}\right) \tag{2}$$

where τ_i is the detection time for the i-th cycle, t_q is the DA signal arrival time for the q-th cycle, and Δt_{jk} represents the delay across specific optical paths (see detailed timing relationship for loose time control in Discussion part).

Figure 1c illustrates the wavelength relay unit (WRU), which leverages efficient on-chip optical-electrical-optical conversion. For broad applicability, we employ a differentially driven add-drop micro-ring modulator (MRM) as the core component. The driving mode and modulator type can be adjusted to suit specific applications, offering flexibility in design and implementation [26]. In the WRU, an optical signal at one wavelength is converted into photocurrent by a photodetector, which then drives a subsequent MRM using supply light at a different wavelength. The operation of the WRU is described by the equation:

$$P_o = f(P_+ - P_-)P_s, \tag{3}$$

where P_+ and P_- denote the optical power of the input light (λ_{n-1}) at the positive and negative ports of the differential photodetectors, respectively, P_s is the supply light power (λ_n), and P_o represents the output power of the MRM. The response function f encapsulates the WRU's operational characteristics. The WRU remains in standby mode until light at λ_{n-1} arrives, triggering its operation. It then stabilizes after completing the current cycle, eliminating strict temporal requirements and thereby ensuring synchronization-free operation. Fig. 1d depicts the electric current flow in the WRU. When the injected current $I_n = I_1 - I_2 + I_b > 0$, the MRM operates in a forward-biased state, resulting in a significant blue shift in its transmission spectrum as I_n increases. Conversely, when $I_n < 0$, the MRM is reversely biased, causing only a minor red shift due to the lower carrier-depletion modulation efficiency [26]. As shown in Fig. 1e, the output light intensity from the through-port of the MRM follows a linearized sigmoid function relative to the driving current. This behavior allows the WRU to perform both linear and nonlinear operations, depending on the MRM's working region. The basic computational tasks, including multiplication, addition, subtraction, and nonlinear transformations, can be executed using the WRU. In synchronous processors, the bandwidth of nonlinear operations must significantly exceed the frequency of data signals to prevent distortion (see Supplementary information S1). As a result, latency in optical links is often redundant under such bandwidth constraints. In contrast, in the asynchronous processor, bandwidth influences only the latency without introducing signal errors, as no serial time signals are processed by the nonlinear unit; the focus is solely on the final steady state after all signals from previous cycles have arrived. To validate the versatility and unique advantages of this architecture, we present two implementations of on-chip optical recurrent accelerators, utilizing the linear and nonlinear regions of the WRU, respectively, as detailed in the subsequent sections.

2.2 Optical hidden Markov model

HMMs are versatile statistical tools widely utilized across diverse domains, including speech recognition, bioinformatics, finance, and natural language processing [27–29]. Figure 2a illustrates the graphical structure of this model, which is composed of the hidden state set $S = \{S_1, S_2, \ldots, S_m\}$ and the observation state set $O = \{O_1, O_2, \ldots, O_n\}$, where m and n represent the number of possible states. The hidden states follow a Markov chain, beginning with an initial state $s_1 \in S$ selected based on a probability vector π and transitioning through subsequent states according to a state

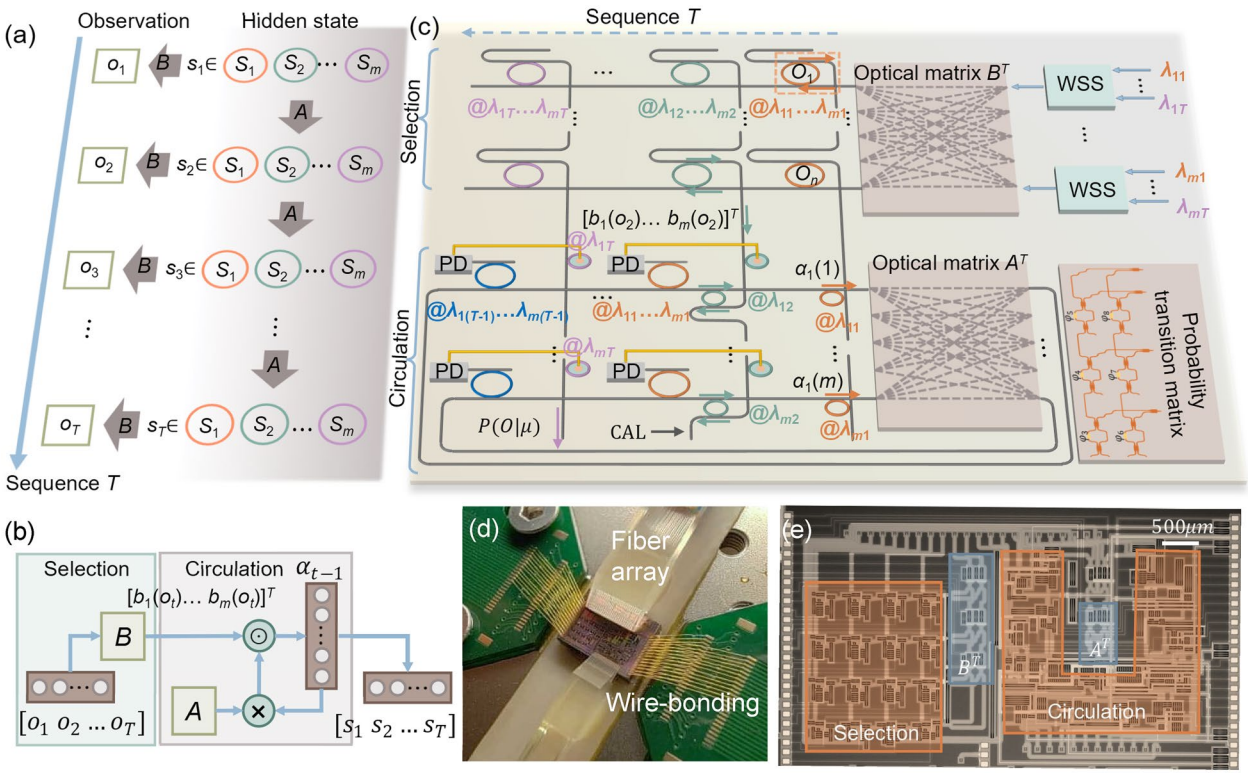

Fig. 2 Principle of the asynchronous OHMM accelerator. **a** Graph structure of the HMM. The right section depicts the possible set of hidden states $S=\{S_1, S_2, \ldots, S_m\}$, while the left section represents the known sequence of observed states $o_1, \ldots, o_T$. Matrix A denotes the transition probability matrix between hidden states, and matrix B corresponds to the observation probability matrix, linking hidden states to observed states. **b** Basic computational flowchart of the HMM, highlighting iterative matrix multiplications involving matrices A and B across multiple cycles. **c** Conceptual diagram of the monolithically integrated OHMM chip. Different wavelengths are used to represent hidden states and are selected via wavelength selective switches (WSS) in the decoding problem. The upper section represents the selection area, while the lower section denotes the circulation area. Probability matrix is realized with the MZI-assisted crossbar array. The processes of multiplication and wavelength relay are implemented through PD-driven MRMs. CAL marks the port for the calibration of computing accuracy. The output intensity from the last cycle represents the probability of the observation sequence. **d** Photograph of the packaged OHMM chip. **e** Layout of the OHMM chip

transition matrix A. After t iterations, a sequence s_1, s_2, ..., s_t is formed. While this sequence remains hidden, an observation state $o_t \in O$ is emitted at each time step t based on the observation probability matrix B. This forms the HMM, formally denoted as (details provided in Supplementary Information S2):

$$\mu = [\pi, A, B] \tag{4}$$

Given an HMM, two fundamental problems arise: (1) evaluating the probability of a specific observation sequence (evaluation problem) and (2) identifying the most probable hidden state sequence corresponding to a given HMM and observation sequence (decoding problem). The core algorithm addressing these problems, as illustrated in Fig. 2b, comprises two primary processes: selection and circulation. It employs the forward variable $\alpha_t(i) = P(o_1 o_2, \ldots, o_t, s_t = S_i | \mu)$, representing the joint probability of a hidden state S_i and an observation sequence $o_1 o_2, \ldots, o_t$ at time t. Using $\alpha_t(i)$, the iterative equation for the forward algorithm (derived in Supplementary Information S3) is expressed as:

$$\begin{bmatrix} \alpha_t(1) \\ \vdots \\ \alpha_t(m) \end{bmatrix} = \begin{cases} \pi \odot \begin{bmatrix} b_1(o_1) \\ \vdots \\ b_m(o_1) \end{bmatrix}, t = 1 \\ \left(A^T \begin{bmatrix} \alpha_{t-1}(1) \\ \vdots \\ \alpha_{t-1}(m) \end{bmatrix} \right) \odot \begin{bmatrix} b_1(o_t) \\ \vdots \\ b_m(o_t) \end{bmatrix}, t > 1 \end{cases} \tag{5}$$

Here, A^T is the transpose of the transition matrix A, $b_i(o_t)$ denotes the observation probability for state S_i, and $\odot$ represents the Hadamard product. This computation involves repeated matrix multiplications, with outputs determined by the observation sequence.

Leveraging the asynchronous recurrent architecture, we present the first implementation of HMM on optical

hardware (OHMM), as shown in Fig. 2c. Wavelength selection switches (WSSs) are used to filter wavelengths corresponding to different hidden states for the decoding problem. In the evaluation problem, all wavelengths enter the chip. During the first cycle, distinct wavelengths (λ_{11}, λ_{21}, ..., λ_{m1}) represent different hidden states, with their intensities indicating the initial probabilities (π) of the corresponding states. Probability matrices are realized using an MZI-assisted crossbar array. In the selection area, the wavelength routing capabilities of the micro-ring resonator (MRR) facilitate the choice of wavelengths corresponding to specific observation states. For example, during $t=1$, the wavelengths λ_{11}, λ_{21}, ..., λ_{m1} undergo multiplication with the optical matrix B^T. Depending on the observation state (e.g., O_1) at that time, a specific large MRR (outlined with dotted lines in Fig. 2c) is tuned to match its resonance peak with these wavelengths, producing $[\alpha_1(1), ..., \alpha_1(m)]$ at the drop port. The large MRRs are characterized by a smaller free spectral range (FSR) and employed to simultaneously route wavelengths associated with all hidden states. Subsequently, in the circulation area, these wavelength signals are dropped by the corresponding small MRRs and undergo further multiplication with the optical matrix A^T. The small MRRs feature a larger FSR and are utilized to route wavelengths corresponding to individual hidden states. The resulting product, $A^T[\alpha_1(1), ..., \alpha_1(m)]^T$, is directed to photodetectors (PDs) through the same large MRRs. The photodetectors convert the optical signals into electrical currents, which drive the MRMs. In the second cycle, the MRMs receive input light intensities represented as $[b_1(o_2), ..., b_m(o_2)]^T$ where the input intensities of λ_{12}, λ_{22}, ..., λ_{m2} to matrix B^T are the same. When operating within their linear region, the MRMs perform the required multiplication operation. This process completes one cycle of Eq. (5). Subsequently, different wavelengths (λ_{i1}, λ_{i2}, ..., λ_{iT} $i=1, ..., m$) are used to represent various cycles. Wavelength relays between adjacent cycles enable the recurrent operations required for the algorithm's progression. The output intensity from the last cycle represents the probability $P(\overline{O}|\mu)$ of a specific observation sequence $\overline{O}$ in the evaluation problem. The decoding problem adopts the same optical hardware as evaluation problem except for different algorithm (see Supplementary Information S3).

Figure 2d presents the packaged photograph of the OHMM chip, fabricated on a standard silicon-on-insulator (SOI) wafer. As a proof of principle, the chip is designed to perform calculations over four cycles, incorporating eight wavelengths in total. A detailed photomicrograph of the chip is shown in Fig. 2e, highlighting its structural components. The chip includes two MZI arrays, which are configured to implement the matrices A^T (2×2) and B^T (4×2) respectively (additional details are provided in Supplementary Information S4). Prior to the experiments, the MRRs were calibrated (characterization of MRRs are provided in Supplementary Information S5), with particular focus on determining the linear operating range of the MRMs. Figure 3a illustrates the linear operating range of the WRU in the chip, showing an extinction ratio of approximately 9 dB, which satisfies the computational requirements. Subsequently, the phase shifters of the MZIs in optical matrices A and B were scanned and adjusted to match the desired matrices. To evaluate the computational accuracy, 500 sets of random vector inputs $[\alpha_1(1)\ \alpha_1(2)]^T$ were used to compute $\sum(A^T[\alpha_1(1)\ \alpha_1(2)]^T)\odot[b_1(o_2)\ b_2(o_2)]^T$ for four different observation states ($o_2 \in\{O_1, O_2, O_3, O_4\}$). Small MRRs corresponding to $t=2$ were initially adjusted to the "through" state, allowing the first cycle's results to be directly output and summed via the CAL port indicated in Fig. 2c, excluding these results from subsequent recursions. By selecting different large MRRs for observation states O_1, O_2, O_3, and O_4, the four elements of the output vector were obtained. The histogram of the correlation coefficient (defined as $corr(\alpha, \beta) = \frac{\alpha\cdot\beta}{|\alpha\cdot\beta|}$) between the theoretical and the experimental output results is presented in Fig. 3b. It demonstrates a high average computational accuracy of 0.9993, with the inset showcasing an example result with a correlation coefficient of 0.9998. Following this, the relay-computation latency of the WRU was measured, as shown in Fig. 3c. The WRU reveals a pulse broadening to approximately 1.76 ns compared with the reference pulse, corresponding to a bandwidth of about 90 MHz (test setup is provided in Supplementary Information S5).

With preparations complete, DNA sequence analysis experiments were conducted to assess the chip's processing capability. Since gene maps can be modeled as HMMs, the two core HMM problems previously described can be analogized to scoring sequences against gene maps and determining the optimal sequence-to-gene map alignment. For this analysis, the yeast mitochondrial gene sequences HS416 and HS3324 were selected [30, 31]. These high-inhibition p-genomes contain a highly similar region, believed to represent the primary origin of wild-type mitochondrial DNA replication. Given the lack of coding regions in these sequences, a two-state binary model was applied to compute smoothed estimates of AT-rich and GC-rich states. As shown in Fig. 3d, the parameters used in the OHMM model are $\pi = \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix}, A = \begin{bmatrix} 0.99 & 0.01 \\ 0.1 & 0.9 \end{bmatrix}, B = \begin{bmatrix} 0.4 & 0.4 & 0.1 & 0.1 \\ 0.05 & 0.05 & 0.4 & 0.5 \end{bmatrix}$ [32 , 33]. As the chip was designed to perform four recursions, the fourth output was feedback-modulated into the input light to facilitate long-sequence processing. The profile plots of the high-inhibition sequences, shown in

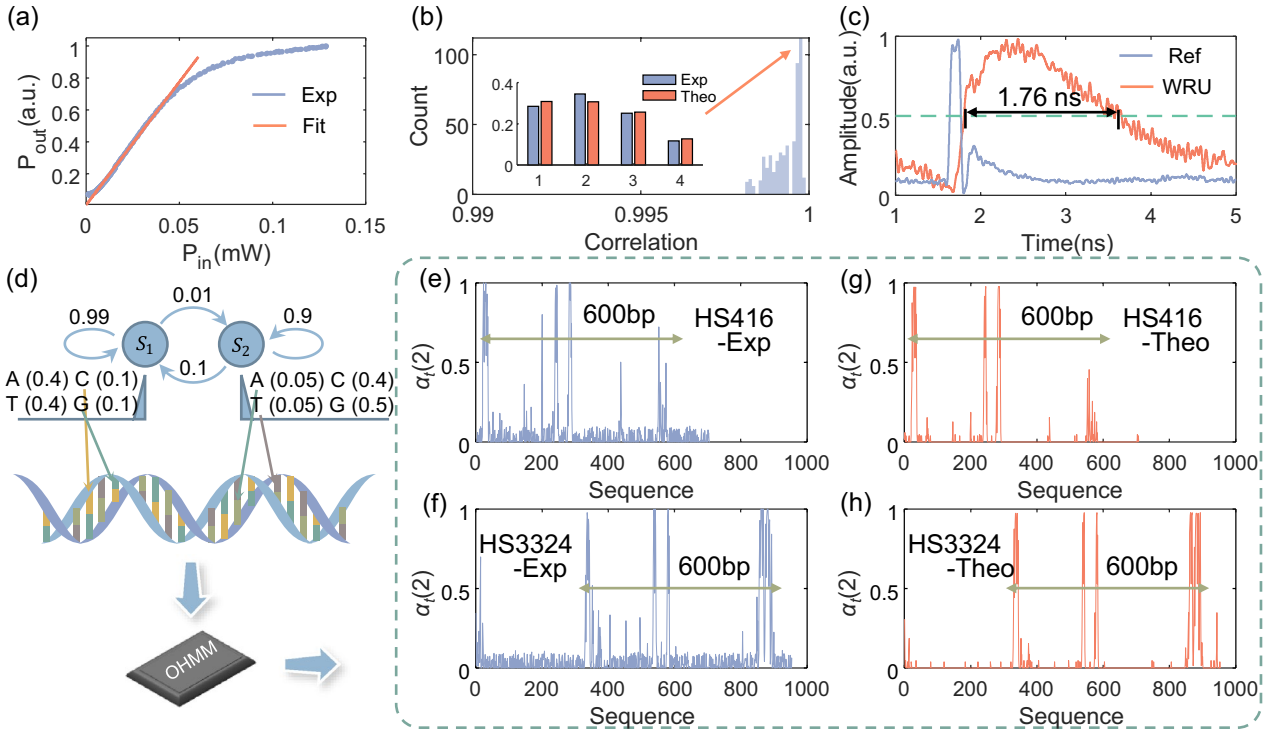

Fig. 3 Experimental results of OHMM. **a** The response curve of WRU within the OHMM chip. The orange line highlights the linear region utilized for recurrent calculations during the experiment. Exp denotes the experimental results. **b** Histogram of correlation coefficients for 500 computed results of a single cycle. The inset illustrates a specific instance with a correlation coefficient of 0.9998. Theo represents the theoretical results. **c** Evaluation of the relay-computation latency (RC-latency) of the WRU. The output pulse from the WRU is broadened to 1.76 ns compared to the reference input pulse. Ref indicates the reference input pulse. **d** A two-state HMM representing DNA sequences with heterogeneous base compositions. State S_1 generates an AT-rich sequence, while state S_2 generates a GC-rich sequence. Arrows indicate state transitions along with their associated probabilities, and the observation probabilities for A, C, G, and T for each hidden state are shown below the respective states. **e**, **f** Calculated sequences of $\alpha_t(2)$ for HS416 and HS3324 using the OHMM chip. **g**, **h** Theoretical sequences of $\alpha_t(2)$ for HS416 and HS3324. Regions where $\alpha_t(2) > 0.5$ correspond to GC-rich regions

Fig. 3e, f, highlight the structural similarity across a region of 600 base pairs (bp). When $\alpha_t(2) > 0.5$, it indicates that the current base pairs are GC-rich; otherwise, they are AT-rich. The experimental results closely align with theoretical outputs (Fig. 3g, h), achieving sequence analysis accuracies of 99.43% and 98.53%, respectively (representing the proportion of matching base pairs between experimental and theoretical data). The influence of nonlinearity on the performance of OHMM is studied in Supplementary Information S6. Additional experimental results, including details on the Chinese word segmentation application, are provided in Supplementary Information S7.

2.3 Optical recurrent neural network

RNNs are a class of artificial neural networks designed to process sequential data. Unlike traditional feedforward neural networks, RNNs incorporate feedback connections, forming directed cycles that enable dynamic temporal behavior. RNNs have found applications across a wide range of domains, including natural language processing, speech recognition, and time series prediction [34]. Figure 4a illustrates the working flow of a standard RNN, which comprises an input layer, a hidden layer, and an output layer. The relationship between the input and hidden layers is described by the following equation:

$$\vec{x}(t) = f_{nl}\left(W_{in}\vec{u}(t) + W\vec{x}(t-\tau)\right), \tag{6}$$

where $\vec{u}(t)$ represents the time-varying input vector, $\vec{x}(t)$ denotes the time-varying hidden vector, W_{in} is the weight matrix for the input vector, W is the feedback weight matrix for the hidden vector, f_{nl} is the nonlinear activation function, and τ is the recurrence period. The relationship between the hidden and output layers is given by:

$$\vec{y}(t) = W_{out}\vec{x}(t), \tag{7}$$

where $\vec{y}(t)$ is the time-varying output vector and W_{out} is the weight matrix for the output. For simplicity, some less critical components in a standard RNN such as the bias vector and the nonlinear activation function for

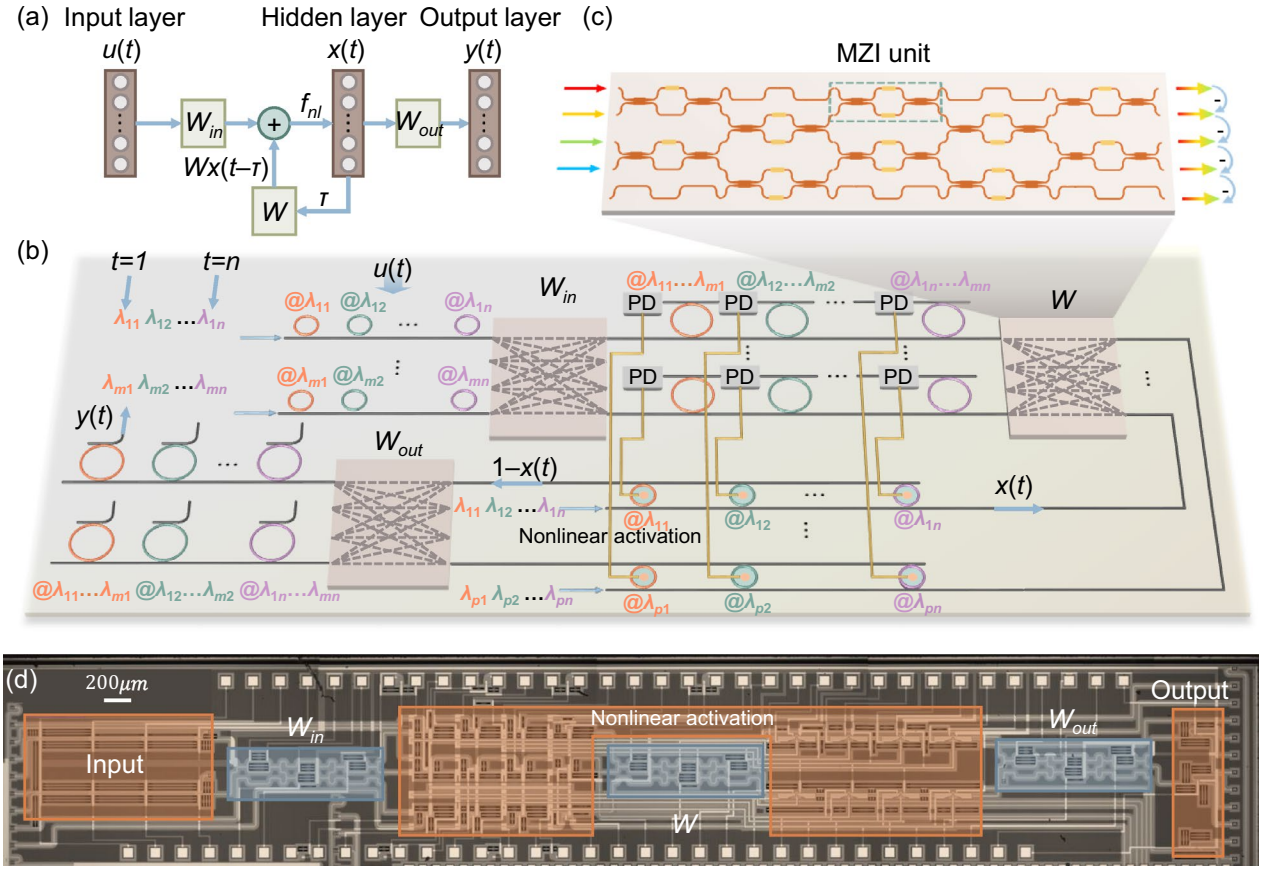

Fig. 4 Principle of the asynchronous ORNN accelerator. **a** The basic working flow diagram of a typical RNN. The network consists of three main layers: the input layer, the hidden layer, and the output layer. The current state of the hidden layer is computed based on the current input layer and the previous state of the hidden layer. **b** The proposed monolithically integrated scheme for the asynchronous ORNN accelerator. The design leverages MRMs driven by PDs to implement both the nonlinear activation function and wavelength relay. The network processes input sequences encoded in different wavelengths asynchronously. **c** The architecture of the optical matrix used in the network. Input signals at different wavelengths are processed by an on-chip incoherent MZI mesh. Real-valued outputs are obtained through differential optical intensity detection. **d** The micrograph of the fabricated ORNN chip

the output layer are omitted. When the dimension of both the input and output layer are reduced to one, the RNN architecture simplifies to a form resembling reservoir computing. Optical reservoir computing has demonstrated significant advancements due to its simplicity, single input/output sequence, and lack of complex synchronization requirements [14, 35]. However, this approach is limited in its ability to handle high-dimensional time series, a challenge that optical RNNs (ORNNs) are better suited to address, making them a subject of deeper investigation [36].

We propose an asynchronous ORNN accelerator, as depicted in Fig. 4b. Here, n sets of wavelengths represent the length of the input time sequences, while the m parallel series correspond to m different wavelengths within the same color. Input information is asynchronously loaded using MRRs and injected into the on-chip incoherent MZI mesh (W_{in}). Large MRRs drop the wavelengths corresponding to the current time step and forward them to photodetectors, which drive the MRM to implement nonlinear activation. Simultaneously, the hidden vector $x(t)$ is fed into the on-chip incoherent MZI mesh (W) in reverse and the wavelengths from the previous cycle will be dropped to the photodetector representing the current cycle. For example, during the second cycle, input signals encoded as green wavelengths are dropped by the green large MRRs, while the hidden signals from the first cycle, encoded as orange wavelengths, are dropped by the orange large MRRs. The drop ports of the respective MRRs are connected to a photodetector, which sums the optical signal intensities and drives the next hidden signal step (green wavelengths). This process completes the calculation of Eq. (6). At the drop port of the MRMs, the operation 1-$x(t)$ is performed and directed into a third on-chip incoherent MZI mesh (W_{out}). The output results y(t) for different time steps are

asynchronously detected after being dropped by corresponding large MRRs. The on-chip incoherent MZI mesh is designed as a simplified real-valued optical matrix, as proposed in our previous work [37] (Fig. 4c; details provided Supplementary Information S8) and the MRMs are driven by differential photocurrents.

The first monolithically integrated ORNN chip was fabricated using the same process as the OHMM chip. As illustrated in Fig. 4d, the dimensions of the input, hidden, and output layers are all two. The chip supports a sequence length of four and is multiplexed by feeding the intermediate results of the fourth cycle back to the first cycle to handle tasks involving longer sequences. To evaluate the chip's performance, a classification task of Japanese vowels was conducted [38]. The original dataset was generated through 12-degree linear prediction analysis, producing a discrete-time series with 12 linear prediction coding (LPC) cepstrum coefficients. To align the dataset dimensions with the chip architecture and meet the positive input requirement, we applied a pre-trained linear dimensionality reduction step followed by a ReLU function. The resulting two-dimensional sequences were used as input for the chip. The classification results were derived from the final cycle of the output sequence (Fig. 5a). The packaged ORNN chip is shown in Fig. 5b. First, the on-chip nonlinear activation function was characterized using differential input optical power, as shown in Fig. 5c, and the results aligned with our theoretical expectations. Following the calibration of the MRRs to their respective working

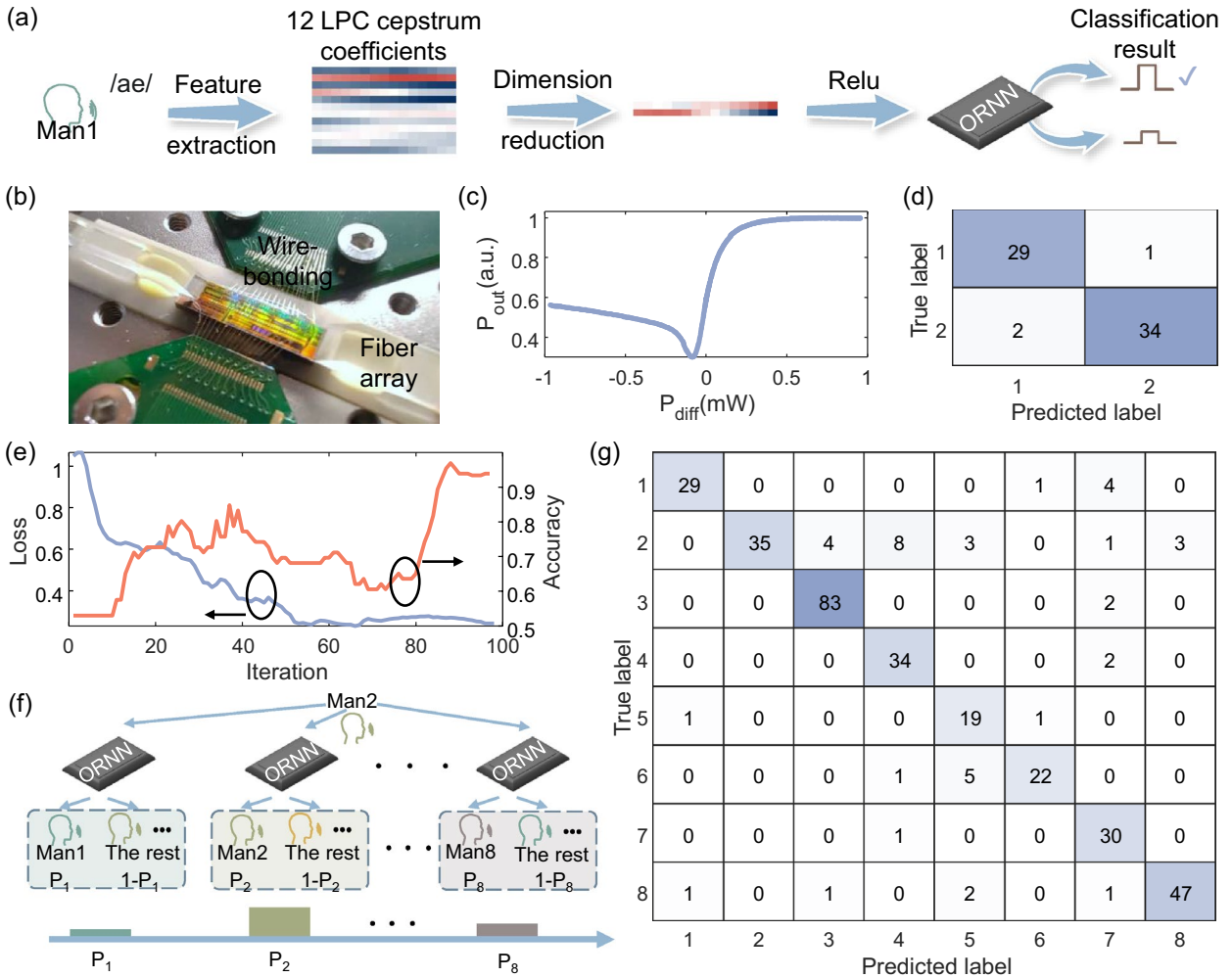

Fig. 5 Experimental demonstration of ORNN. **a** Workflow of the audio classification task. The 12-dimensional sequential input signals are first reduced to two dimensions through a linear transformation followed by a ReLU function. These reduced signals are then processed by the ORNN chip, with the final output signal intensities determining the classification result. **b** Photograph of the packaged ORNN chip. **c** Measured nonlinear activation function of the differential PD-driven MRM. Here, P_{diff} and P_{out} represent the input differential optical power and the output optical power after nonlinear activation, respectively. **d** Confusion matrix for the test dataset in the binary classification task. **e** Iterative curves showing the loss function and accuracy progression in the binary classification task. **f** Schematic flow diagram illustrating the one-versus-rest method employed during the training of the eight-class classification task. **g** Confusion matrix for the test dataset in the eight-class classification task

wavelengths, the chip was operational. The softmax function was applied to calculate the probability distribution for each class, and entropy loss was selected as the loss function, which was minimized in-situ using a stochastic parallel gradient descent algorithm (see "Methods" section). The two-classification achieved accuracies of 97% and 95% for the training and test datasets, respectively, with the corresponding confusion matrix depicted in Fig. 5d. The progression of the loss function and accuracy during training is presented in Fig. 5e. To further demonstrate the processing capability of the ORNN chip, we extended its application to an eight-class classification task. Addressing multi-class problems requires an appropriate classifier. In this work, the one-versus-rest method was employed [39]. This approach utilized eight parallel ORNNs to classify speech sequences into eight categories, as illustrated in Fig. 5f. During training, each category was treated as a separate class, with the remaining categories grouped as another. When presented with an unknown speech sequence, the entire architecture outputs eight classification probabilities, with the predicted label corresponding to the class with the highest probability. After in-situ training, the confusion matrix for the eight-class classification is shown in Fig. 5g, demonstrating a test accuracy of 87.7% (details on the training process are provided in Supplementary Information S9).

3 Discussion

3.1 Comparison with other asynchronous computing

In this section, we compare our scheme with traditional asynchronous computing by conducting a detailed analysis of the timing relationships within the proposed architecture (Fig. 6). Asynchronous computing is a widely adopted concept in electronic circuits and programs, where tasks or operations are executed independently, without requiring the completion of other tasks before starting [40]. Fig. 6 depicts the input DA signals and corresponding outputs from each cycle. For instance, DAC_1 generates an input signal, and its output for each cycle is a temporally misaligned superposition of multiple signals ($Output_1$) due to distinct latency in each computational link. Consequently, the output of each cycle encompasses computing results from all possible links associated with different DACs. In this context, accurate computation requires signals from all links to temporally overlap. The margin time for asynchronous computing accounts for the ADC sampling time, time misalignment among different DA signals, and latency misalignment across various signal paths. Among these, the ADC sampling time and time misalignment among DA signals constitute a constant value, denoted as τ_0. Consequently, the duration of the q-th DA signal, $\Delta\tau_q$, can be expressed as:

$$\Delta\tau_q > \tau_0 + \sum_{k=q}^{s} \max_j \Delta t_{jk} - \sum_{k=q}^{s} \min_j \Delta t_{jk}. \tag{8}$$

This condition ensures that each link, from input to output, operates as an independent task, necessitating only loose time control to complete all tasks within the ADC sampling span. Overall, the core principle of our proposed scheme shares similarities with traditional asynchronous computing. However, our asynchronous approach is relatively constrained, as signals must be sampled within a specific time span. As we will discuss in the following analysis, this loose time control mechanism offers a significant enhancement in computational efficiency compared to completely time-independent

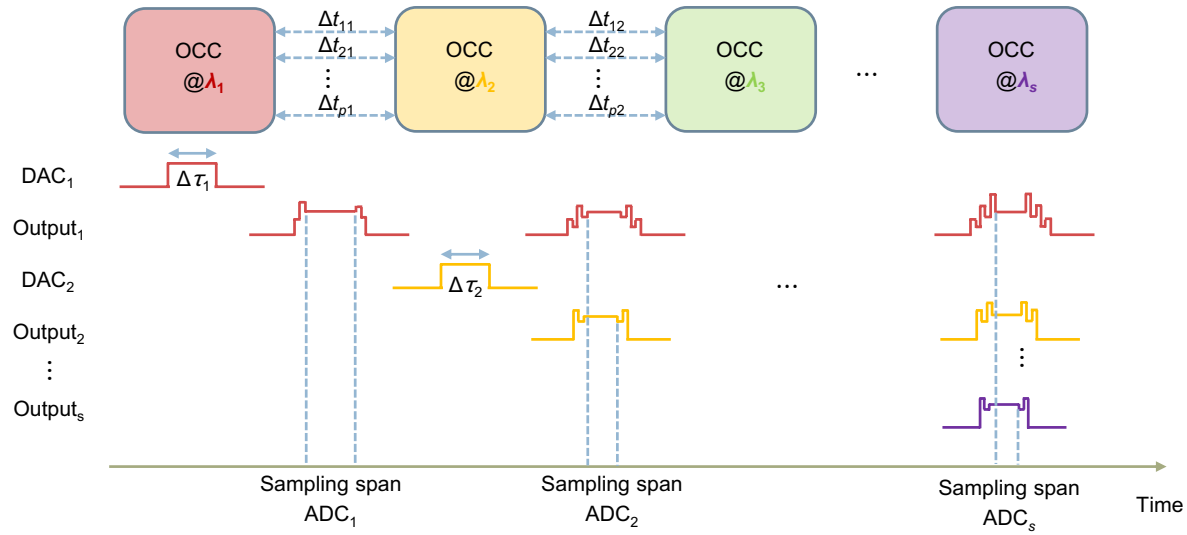

Fig. 6 Timing relationships in our proposed optical asynchronous recurrent accelerator

operations ($\Delta\tau_q$ is equal to or longer than the total latency, Fig. 1b).

3.2 Management and expansion of wavelength resources

The scalability of our scheme is primarily determined by the number of available wavelengths. Significant advancements in on-chip multi-wavelength optical sources have enabled the generation of up to 200 wavelengths in the telecom band [41]. Another critical component in our setup is the wavelength demultiplexer. Due to the limited free spectral range (FSR) of MRRs, they can typically support a maximum of 30 wavelengths [42]. To address this limitation, a potential solution involves leveraging cascaded optical interleavers and MRRs to manage massive wavelength Kerr combs [43]. Additionally, nanobeam cavities, which are not constrained by FSR, offer a promising alternative for demultiplexing large numbers of wavelengths [44]. Current state-of-the-art wavelength division multiplexers can accommodate up to 512 wavelengths with 10 GHz spacing [45], a capability approaching the upper limit of existing comb sources. While the proposed asynchronous architecture can manage the sequence lengths required for most recurrent tasks, it is also capable of time-multiplexing to accommodate applications that demand ultra-long sequence processing. This is achieved by feeding the intermediate results of the final cycle in the first s sequences back to the initial cycle of the second s sequences. While this reuse of the chip introduces additional power consumption and latency in the electrical domain, the averaged overhead per cycle becomes negligible when s is sufficiently large.

3.3 Overhead analysis of wavelength-relay unit

In our demonstration, the power-efficient WRU is employed to achieve wavelength relays. Notably, the WRU also performs a critical role as a nonlinear function in many recurrent applications [46], without which synchronous optical processors would be restricted to simple linear recursion. The power consumption of the WRU is determined by the signal light power and the electrical power consumed by the photodetector, as expressed by:

$$P_{total} = P_{opt} + P_{elect} = \frac{I_{drive}}{\eta_{PD}} + V_{bias} I_{drive}, \tag{9}$$

where I_{drive} is the driving current of the modulator, η_{PD} is the responsivity of the photodetector, and V_{bias} is the bias voltage of the photodetector. In our experiment, these parameters are I_{drive}=0.09 mA, η_{PD}=0.9 A/W, and V_{bias}=3 V, resulting in a power consumption of 0.37 mW. These values can be optimized to further reduce power consumption. For instance, the power consumption of the state-of-the-art WRU can be minimized to 0.15 mW [47]. To further minimize the energy consumption of the WRU, its operational duration must be carefully optimized. As illustrated in Fig. 6, the WRU's working duration closely correlates with the duration of the DA signal. Consequently, $\Delta\tau_q$ can be selected as the minimum value that satisfies the condition in Eq. (8). The latency of the WRU is mainly influenced by the RC response time, which is relatively large due to the use of standard library devices without specific design optimizations. Nevertheless, with targeted engineering, the WRU's bandwidth can exceed 1 GHz, reducing the corresponding latency to below 100 ps [47].

3.4 Key performance of different optical recurrent computing architectures

We comprehensively analyze the latency and energy efficiency of both synchronous and asynchronous architectures in Supplementary Information S10. Considering the latencies of the optical link and WRU, the total latencies for one cycle of the OHMM and ORNN are 1.83 ns and 1.82 ns, respectively. These values are six orders of magnitude lower than those of the previously reported spatial ORNN systems [36]. Using parameters from the literature [20], the energy efficiency of the ORNN chip is calculated to be 0.48 TOPs/J. This value can be further improved by scaling up the computational workload. For instance, assuming a computing scale of 64, the estimated energy efficiency increases to 11.62 TOPs/J, which is an order of magnitude higher than that of the spatial ORNN [36].

Table 1 provides a comparison of key features—latency and energy consumption—among three architectures. For the electronic feedback scheme, the latency and energy efficiency are estimated using parameters from a typical FPGA [48], with a transmission latency of 0.5 μs and an energy efficiency of 28.2 pJ/OP. The asynchronous architecture achieves nearly four times the energy efficiency of the synchronous architecture. This advantage arises because the sampling rate (F) of the synchronous architecture is four times the data rate (B) when operating with a sequence length of four (see Supplementary Information S10). It can be further enhanced by involving more cycles on the same chip. As shown in Fig. 7a, although the asynchronous architecture demonstrates superior energy efficiency compared to the synchronous and electrical feedback architectures, its energy efficiency decreases with the sequence length due to the quadratic growth in the energy consumption of WRUs. This drawback can be mitigated by reducing $\Delta\tau_q$ or lowering the power consumption of individual WRUs. The advantage of the asynchronous architecture over the electrical

Table 1 Comparison of different optical recurrent architectures on the main performance (s is the number of computing cycles)

Architecture	Latency	Energy efficiency	Delicate delay control	Signal distortion
Synchronous architecture (2×2, $s=4$)	7.28 ns	0.12 TOPs/J ($F/B=4$)	Yes	Yes
Synchronous architecture (64×64, $s=12$)	21.84 ns	0.99 TOPs/J ($F/B=12$)	Yes	Yes
Electrical feedback [48] (2×2, $s=4$)	2 μs	0.08 TOPs/J	No	No
Electrical feedback [48] (64×64, $s=12$)	6 μs	1.96 TOPs/J	No	No
Spatial ORNN [36] (490,000×490,000, $s=1$)	8 ms	1.58 TOPs/J	No	No
Our work (2×2, $s=4$)	7.28 ns	0.48 TOPs/J	No	No
Our work (64×64, $s=12$)	21.84 ns	11.62 TOPs/J	No	No

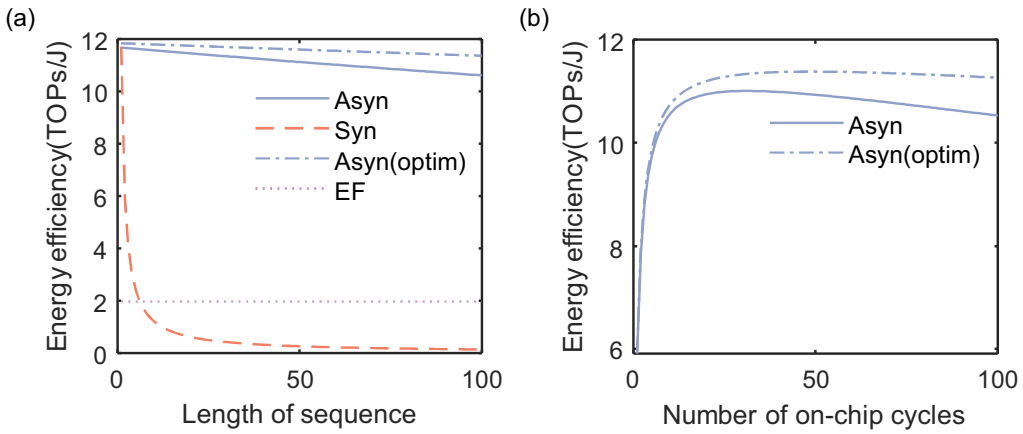

Fig. 7 **a** Energy efficiency comparison of various optical recurrent architectures as a function of sequence length for a matrix scale of 64×64. "Asyn" represents the asynchronous architecture, "Syn" refers to the synchronous architecture, "Optim" corresponds to the asynchronous architecture optimized with a WRU power consumption of 0.15 mW, and "EF" denotes the electrical feedback architecture. **b** Energy efficiency of the asynchronous architecture as a function of the number of on-chip cycles in the chip-reuse strategy

feedback scheme can be further amplified by incorporating more complex recurrent operations, such as increasing the number of recurrent vectors in the RNN model. To maximize energy efficiency, the optimal number of on-chip cycles can be determined, and a chip-reuse strategy can be employed to efficiently process long sequence signals (Fig. 7b). In the future, the scalability of our method for large-scale optical computing requires further advancements in integration density, manufacturing technologies, and the energy efficiency of core optical components, such as the MZI mesh, wavelength multiplexers, and WRUs. Beyond intuitive metrics like latency and energy consumption, the proposed asynchronous architecture eliminates the need for delicate delay control and mitigates signal distortion inherent in synchronous architectures. These factors, including energy consumption and latency induced by precise delay controls, as well as signal degradation impairing computing accuracy, are significant challenges in synchronous designs.

4 Conclusions

In conclusion, we present a novel asynchronous architecture for an on-chip optical recurrent accelerator, leveraging time-wavelength mapping and wavelength relay. This architecture effectively overcomes the critical synchronization challenges that hinder the implementation of parallel synchronous optical recurrent accelerators. Through detailed analysis, we highlight the proposed scheme's advantages in achieving low latency and energy consumption. Additionally, we demonstrate two monolithically integrated prototypes: the OHMM and ORNN. Their exceptional performance on various benchmark tasks underscores the versatility and potential of the asynchronous recurrent accelerator. This proposed architecture paves a practical pathway for large-scale parallel sequential signal processing using photonic hardware, with promising applications in domains such as autonomous driving and intelligent robotics.

5 Methods

5.1 Stochastic parallel gradient descent algorithm

In the in-situ training of the ORNN, we employ the stochastic parallel gradient descent algorithm to estimate the gradient of the loss function [49]. In each iteration, a random perturbation vector δ is generated and applied to the current voltages as $U+\delta$ and $U-\delta$. The corresponding loss function values, $L(U+\delta)$ and $L(U-\delta)$, are then computed. The estimated gradient of the loss function is given by:

$$G = 2\delta[L(U+\delta) - L(U-\delta)]. \tag{10}$$

Subsequently, the voltages are updated using the Adam algorithm, a fast-converging gradient descent optimization method [50]

$$\begin{aligned} U(iter+1) =& U(iter) + \alpha\left(v_{iter}/(1-\beta_1^{iter})\right)/\sqrt{s_{iter}/(1-\beta_2^{iter})+\varepsilon}, \\ v_{iter} =& \beta_1 v_{iter-1} + (1-\beta_1)G, \\ s_{iter} =& \beta_2 s_{iter-1} + (1-\beta_2)G^2, \end{aligned} \tag{11}$$

where *iter* is the current iteration, α is the learning rate (set to 0.1 during training), and β_1, β_2, ε are hyperparameters with values 0.9, 0.999, and 10^{-8}, respectively. The initial values of v_{iter} and s_{iter} are set to zero.

5.2 Experimental methods

The chip was fabricated using a 200 mm CMOS process line with a two-layer copper interconnect. The fabrication line width was as narrow as 130 nm, achieved through a deep ultraviolet lithography process. The on-chip photodetector utilizes a lateral PIN structure with an epitaxial germanium layer of 260 nm thickness. The chip's optical I/O consists of a vertical grating coupler array packaged with a horizontally coupled fiber array, while the electrical I/O is connected to the PCB via wire bonding. Calibration of the MRRs and MRMs on the chip was performed by sweeping their thermal phase shifters. For latency measurements, the input laser source was modulated using a lithium niobate intensity modulator with a 10 GHz bandwidth. A bit pattern generator produced the pulse signal, with a pulse width of 100 ps and a period of 4 ns. The output optical signal was detected by a photodetector with an 18 GHz bandwidth and captured by a Tektronix DSA72004B oscilloscope. The thermal phase shifters and the WRU bias voltage were powered by a digital-to-analog converter (LTC2688), which was controlled by a field-programmable gate array (FPGA) chip (7K325T). A personal computer managed the entire experimental system via serial ports. To ensure stability during operation, the entire chip was thermally stabilized using a thermoelectric cooler (TEC).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43593-025-00084-y.

Additional file 1.

Acknowledgements
The work was supported by the Fundamental Research Funds for the Central Universities.

Author contributions
BW, HJZ, JJD, and HLZ conceived the idea. BW and HJZ designed and fabricated the chip. BW and HJZ carried out theoretical analysis and simulation. BW and HJZ designed and performed the experiments. JWC, WKZ, SJZ, HLZ, JJD, BW, and HJZ discussed and analyzed data. BW and HJZ prepared the manuscript. CRH, DMH, HLZ, and JJD revised the paper and XLZ supervised the project. All authors contributed to the writing of the manuscript.

Funding
National Key Research and Development Project of China (2023YFB2806502); National Natural Science Foundation of China (62425504, U21A20511, 62275088, 62075075); Knowledge Innovation Program of Wuhan-Basic Research 2023010201010049.

Availability of data and materials
The datasets used and/or analyzed in the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
There is no ethics issue for this paper.

Consent for publication
All authors agreed to publish this paper.

Competing interests
The authors declare that they have no competing interests.

Received: 27 October 2024 Revised: 17 January 2025 Accepted: 4 March 2025
Published online: 06 May 2025

References

1. H. Zhou et al., Photonic matrix multiplication lights up photonic accelerator and beyond. Light Sci. Appl. **11**, 30 (2022). https://doi.org/10.1038/s41377-022-00717-8
2. C. Li, X. Zhang, J. Li, T. Fang, X. Dong, The challenges of modern computing and new opportunities for optics. PhotoniX (2021). https://doi.org/10.1186/s43074-021-00042-0
3. B. Wu, H. Zhou, J. Dong, X. Zhang, Programmable integrated photonic coherent matrix: principle, configuring, and applications. Appl. Phys. Rev. (2024). https://doi.org/10.1063/5.0184982
4. Y. Chen et al., All-analog photoelectronic chip for high-speed vision tasks. Nature (2023). https://doi.org/10.1038/s41586-023-06558-8
5. T. Wang et al., Image sensing with multilayer nonlinear optical neural networks. Nat. Photonics (2023). https://doi.org/10.1038/s41566-023-01170-8
6. S. Pai et al., Experimental evaluation of digitally verifiable photonic computing for blockchain and cryptocurrency. Optica (2023). https://doi.org/10.1364/optica.476173
7. K. Vandoorne et al., Experimental demonstration of reservoir computing on a silicon photonics chip. Nat. Commun. **5**, 3541 (2014). https://doi.org/10.1038/ncomms4541

8. T. Zhou, W. Wu, J. Zhang, S. Yu, L. Fang, Ultrafast dynamic machine vision with spatiotemporal photonic computing. Sci. Adv. **9**, eadg4391 (2023). https://doi.org/10.1126/sciadv.adg4391
9. T. Honjo et al., 100,000-spin coherent Ising machine. Sci. Adv. **7**, eabh0952 (2021). https://doi.org/10.1126/sciadv.abh0952
10. P.L. McMahon et al., A fully programmable 100-spin coherent Ising machine with all-to-all connections. Science **354**, 614–617 (2016). https://doi.org/10.1126/science.aah5178
11. T. Inagaki et al., A coherent Ising machine for 2000-node optimization problems. Science **354**, 603–606 (2016). https://doi.org/10.1126/science.aah4243
12. X. Liu, J. Cheng, H. Zhou, J. Dong, X. Zhang, Chip-scale all-optical complex-valued matrix inverter. APL Photonics (2024). https://doi.org/10.1063/5.0200149
13. D.C. Tzarouchis, M.J. Mencagli, B. Edwards, N. Engheta, Mathematical operations and equation solving with reconfigurable metadevices. Light Sci. Appl. (2022). https://doi.org/10.1038/s41377-022-00950-1
14. Y.-W. Shen et al., Deep photonic reservoir computing recurrent network. Optica (2023). https://doi.org/10.1364/optica.506635
15. Y. Shen et al., Deep learning with coherent nanophotonic circuits. Nat. Photonics **11**, 441–446 (2017). https://doi.org/10.1038/nphoton.2017.93
16. H. Sun, K.-T. Wu, *Enabling Technologies for High Spectral-Efficiency Coherent Optical Communication Networks* (Wiley, London, 2016), pp.355–394
17. L. Zhao, H. Dai, C. Yang, J. Lu, Y. Sun, Tradeoff between time and energy costs for controlling stochastic coupled neural networks. IEEE Trans. Autom. Control **69**, 1112–1118 (2024). https://doi.org/10.1109/tac.2023.3281471
18. S. Wang, M. Shi, D. Li and T. Du, A Survey of Time Synchronization Algorithms for Wireless Sensor Networks. 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019, pp. 6392–6397. https://doi.org/10.23919/ChiCC.2019.8866385
19. F. Rivet et al., The experimental demonstration of a SASP-based full software radio receiver. IEEE J. Solid State Circ. **45**, 979–988 (2010). https://doi.org/10.1109/jssc.2010.2041402
20. B.S.G. Pillai et al., End-to-end energy modeling and analysis of long-haul coherent transmission systems. J. Lightw. Technol. **32**, 3093–3111 (2014). https://doi.org/10.1109/jlt.2014.2331086
21. M. Bahadori et al., Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing. Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017, Lausanne, Switzerland, 2017, pp. 326–331. https://doi.org/10.23919/DATE.2017.7927010
22. L.I. Gang, Z. Li-jun, L.I.N. Ling, H.E. Feng, Application of oversampling technique in upgrading ADC's resolution in weak signal detection. Nanotechnol. Precis. Eng. **7**, 71–75 (2009)
23. M. Prabhu et al., Accelerating recurrent Ising machines in photonic integrated circuits. Optica (2020). https://doi.org/10.1364/optica.386613
24. J. Cheng et al., Multimodal deep learning using on-chip diffractive optics with in situ training capability. Nat. Commun. **15**, 6189 (2024). https://doi.org/10.1038/s41467-024-50677-3
25. Z. Xu et al., Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. Science **384**, 202–209 (2024). https://doi.org/10.1126/science.adl1203
26. A.N. Tait et al., Silicon photonic modulator neuron. Phys. Rev. Appl. (2019). https://doi.org/10.1103/PhysRevApplied.11.064043
27. B. Mor, S. Garhwal, A. Kumar, A systematic review of hidden Markov models and their applications. Arch. Comput. Methods Eng. **28**, 1429–1448 (2020). https://doi.org/10.1007/s11831-020-09422-4
28. A. Coutrot, J.H. Hsiao, A.B. Chan, Scanpath modeling and classification with hidden Markov models. Behav. Res. Methods **50**, 362–379 (2017). https://doi.org/10.3758/s13428-017-0876-8
29. Y. Khalifa, D. Mandic, E. Sejdić, A review of hidden Markov models and recurrent neural networks for event detection and localization in biomedical signals. Inform. Fusion **69**, 52–72 (2021). https://doi.org/10.1016/j.inffus.2020.11.008
30. H. Blanc, B. Dujon, Replicator regions of the yeast mitochondrial DNA responsible for suppressiveness. Proc. Natl. Acad. Sci. **77**, 3942–3946 (1980). https://doi.org/10.1073/pnas.77.7.3942
31. H. Blanc, Two modules from the hypersuppressive rho—mitochondrial DNA are required for plasmid replication in yeast. Gene **30**, 47–61 (1984). https://doi.org/10.1016/0378-1119(84)90104-5
32. G.A. Churchill, Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. **51**, 79–94 (1989). https://doi.org/10.1016/S0092-8240(89)80049-7
33. S.R. Eddy, Hidden Markov models. Curr. Opin. Struct. Biol. **6**, 361–365 (1996). https://doi.org/10.1016/S0959-440X(96)80056-X
34. S. Das, A. Tariq, T. Santos, S.S. Kantareddy, I. Banerjee, *Machine Learning for Brain Disorders,* ed. by Olivier Colliot (Springer, London, 2023), pp. 117–138
35. A. Lupo, E. Picco, M. Zajnulina, S. Massar, Deep photonic reservoir computer based on frequency multiplexing with fully analog connection between layers. Optica (2023). https://doi.org/10.1364/optica.489501
36. T.K. Zhou et al., Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. Nat. Photonics **15**, 367–373 (2021). https://doi.org/10.1038/s41566-021-00796-w
37. B. Wu et al., Real-valued optical matrix computing with simplified MZI mesh. Intell. Comput. **2**, 0047 (2023). https://doi.org/10.34133/icomputing.0047
38. M. Kudo, J. Toyama, M. Shimbo, *Japanese Vowels Data Set. Distributed by UCI Machine Learning Repository*
39. J. Xu, An extended one-versus-rest support vector machine for multi-label classification. Neurocomputing **74**, 3114–3124 (2011). https://doi.org/10.1016/j.neucom.2011.04.024
40. M. Herlihy, N. Shavit, The topological structure of asynchronous computability. J. ACM **46**, 858–923 (1999). https://doi.org/10.1145/331524.331529
41. H.S. Stokowski et al., Integrated frequency-modulated optical parametric oscillator. Nature **627**, 95–100 (2024). https://doi.org/10.1038/s41586-024-07071-2
42. A. Jha et al., Nanophotonic cavity based synapse for scalable photonic neural networks. IEEE J. Sel. Top. Quantum Electron. **28**, 1–8 (2022). https://doi.org/10.1109/jstqe.2022.3179983
43. A. Rizzo et al., Massively scalable Kerr comb-driven silicon photonic link. Nat. Photonics (2023). https://doi.org/10.1038/s41566-023-01244-7
44. J. Zhang, B. Wu, J. Cheng, J. Dong, X. Zhang, Compact, efficient, and scalable nanobeam core for photonic matrix-vector multiplication. Optica (2024). https://doi.org/10.1364/optica.506603
45. K. Takada, M. Abe, M. Shibata, M. Ishii, K. Okamoto, Low-crosstalk 10-GHz-spaced 512-channel arrayed-waveguide grating multi/demultiplexer fabricated on a 4-in wafer. IEEE Photonics Technol. Lett. **13**, 1182–1184 (2001). https://doi.org/10.1109/68.959357
46. F. Bohm, G. Verschaffelt, G. Van der Sande, A poor man's coherent Ising machine based on opto-electronic feedback systems for solving optimization problems. Nat. Commun. **10**, 3538 (2019). https://doi.org/10.1038/s41467-019-11484-3
47. K. Nozaki et al., Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions. Nat. Photonics **13**, 454–459 (2019). https://doi.org/10.1038/s41566-019-0397-3
48. Z. Li et al., Scalable on-chip optoelectronic ising machine utilizing thin-film lithium niobate photonics. ACS Photonics (2024). https://doi.org/10.1021/acsphotonics.4c00003
49. Y. Wan et al., Efficient stochastic parallel gradient descent training for on-chip optical processor. Opto Electron. Adv. **7**, 230182–230182 (2024). https://doi.org/10.29026/oea.2024.230182
50. D.P. Kingma, J.A. Ba, A method for stochastic optimization. arXiv:1412.6980 (2014).