nature photonics

Article

https://doi.org/10.1038/s41566-023-01233-w

Deep learning with coherent VCSEL neural networks

Received: 24 August 2022

Accepted: 16 May 2023

Published online: 17 July 2023

Check for updates

Zaijun Chen[1,2] ✉, Alexander Sludds[1], Ronald Davis III[1], Ian Christen[1], Liane Bernstein[1], Lamia Ateshian[1], Tobias Heuser[3], Niels Heermeier[3], James A. Lott[3], Stephan Reitzenstein[3], Ryan Hamerly[1,4] ✉ & Dirk Englund[1] ✉

Deep neural networks (DNNs) are reshaping the field of information processing. With the exponential growth of these DNNs challenging existing computing hardware, optical neural networks (ONNs) have recently emerged to process DNN tasks with high clock rates, parallelism and low-loss data transmission. However, existing challenges for ONNs are high energy consumption due to their low electro-optic conversion efficiency, low compute density due to large device footprints and channel crosstalk, and long latency due to the lack of inline nonlinearity. Here we experimentally demonstrate a spatial-temporal-multiplexed ONN system that simultaneously overcomes all these challenges. We exploit neuron encoding with volume-manufactured micrometre-scale vertical-cavity surface-emitting laser (VCSEL) arrays that exhibit efficient electro-optic conversion (<5 attojoules per symbol with a π-phase-shift voltage of $V_\pi = 4$ mV) and compact footprint (<0.01 mm^2 per device). Homodyne photoelectric multiplication allows matrix operations at the quantum-noise limit and detection-based optical nonlinearity with instantaneous response. With three-dimensional neural connectivity, our system can reach an energy efficiency of 7 femtojoules per operation (OP) with a compute density of 6 teraOP mm^{-2} s^{-1}, representing 100-fold and 20-fold improvements, respectively, over state-of-the-art digital processors. Near-term development could improve these metrics by two more orders of magnitude. Our optoelectronic processor opens new avenues to accelerate machine learning tasks from data centres to decentralized devices.

Artificial neural networks are computational systems that imitate the way in which biological brains process information. These systems are built to learn, combine and summarize information from large datasets. As a result of the advances in deep neural network (DNN) algorithms and also increases in computing power, DNNs[1] have thrived in recent years and revolutionized information processing in applications including image[2], object[3] and speech recognition[4], game playing[5], medicine[6] and physical chemistry[7]. A deep fully connected neural network is made of N layers (Fig. 1a), where each layer consists of a matrix-vector multiplication and a nonlinear activation. Driven by the need to tackle problems of increasing complexity, the size of machine learning models is increasing exponentially[8], with some reaching more than 100 billion trainable

[1]Research Laboratory of Electronics, MIT, Cambridge, MA, USA. [2]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. [3]Fakultät II Institut für Festkörperphysik, Technische Universität Berlin, Berlin, Germany. [4]NTT Research Inc., PHI Laboratories, Sunnyvale, CA, USA. ✉e-mail: zaijunch@usc.edu; rhamerly@mit.edu; englund@mit.edu

https://doi.org/10.1038/s41566-023-01233-w

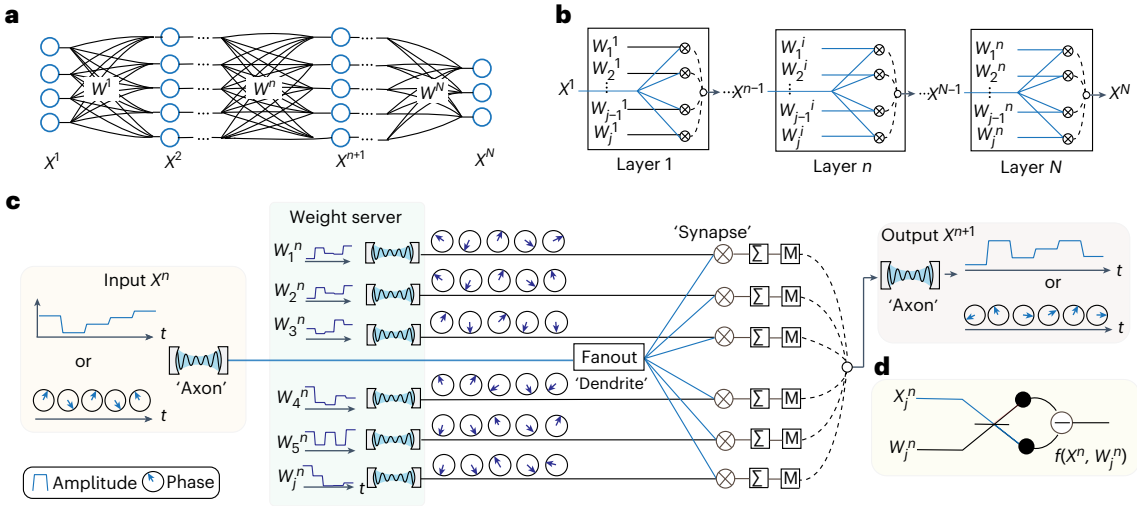

Fig. 1 | VCSEL-ONN architecture. a, Representation of a fully connected DNN composed of *N* layers. **b**, Our implementation of an *N*-layer neural network with an optical tensor processor in each layer. **c**, Detailed illustration of the *n*th ONN layer. The 'axon' laser oscillator encoding the input vector X^n is fanned out, allowing for parallel computing with *j* synaptic weight vectors. The weights are generated from a server of *j* laser oscillators and can potentially be broadcast to process multiple inputs in parallel. Multiply-and-accumulate operations are based on homodyne detection and time integration. The integrated values are stored in nearby digital memories (labelled 'M') and subsequently serialized as the input vector to the next layer. **d**, Homodyne balance detection. Optical interference between two laser fields generates homodyne product $f(X^n, W_j^n)$. ∑, integrator; M, digital memory.

parameters as of 2020 (ref. 9). In contrast, due to the practical limits on transistor counts[10] and energy consumption in data movement[11], extending computational capacity with complementary–metal–oxide–semiconductor (CMOS) circuits has become more and more difficult. An alternative approach leveraging qualitatively different technology must be developed to continue the scaling of computing power in the coming decades.

Several critical bottlenecks emerge when designing efficient and scalable neural network accelerators. Table 1 summarizes the key figures of merit, based on several recent studies[12–14]. State-of-the-art electronic microprocessors, such as graphics processing units (GPUs)[15] and application-specific integrated circuits (ASICs)[16], are optimized for machine learning tasks by means of the energy efficiency (the energy consumption per operation) (C1), $\epsilon \approx 1$ pJ/OP (refs. 16,17) and the compute density (the number of operations per second for a given chip area), (C2) $\rho \approx 0.35$ teraOP mm^{-2} s^{-1} (NVDIA A100 GPU), limited by the wire capacitance of electronic interconnects[18]).

Optical neural networks (ONNs) hold great promise to alleviate this bottleneck, with orders-of-magnitude improvement in criteria C1 and C2 due to their large optical bandwidth and low-loss data transmission[19]. Recent progress in ONNs has demonstrated fully connected layers with photonic integrated circuits[20–24] and holographic phase masks[25,26], linear matrix operations at high throughput[27] and low-energy optical readout[28,29]. However, due to the large footprint and high electro-optic (EO) energy cost of the photonic devices (for example, state-of-the-art low-loss EO modulators require $V_\pi L > 1$ V cm (ref. 30), where *L* is the device length), simultaneously achieving C1 and C2 remains an unfulfilled challenge. Moreover, incorporating low-energy, all-optical nonlinearity into ONNs is challenging due to the weak photon–photon interaction. Recent advances rely on resonant cavities[23], laser-cooled atoms[31] and femtosecond pulses in (millimetre-) long waveguides[32] to enhance the nonlinearity, pointing to a promising potential of speed-of-light activation, but these systems are either slow due to the cavity lifetime or limited coupling strength between atomic levels (slow Rabi oscillations), or bulky due to the device and instrument dimensions. Alternatively, most ONNs implement the nonlinear activation function digitally[20,25,27,29] or optoelectronically[21,22,33], resulting in latency or energy constraints. A fast, compact, and low-energy C3 nonlinearity has yet to be developed. Furthermore, to meet the demanding scaling of DNN models, photonic neuromorphic devices should be scalable (C4), with high density, to extend the computing power while reducing fabrication cost. Furthermore, the ONN architecture should be scalable in the number of neurons to support large DNN models (C5).

In this Article, we introduce a compact VCSEL-ONN architecture that achieves all five criteria (C1–C5) simultaneously. We explore the potential of high-speed VCSELs for next-generation ONNs, especially now that VCSEL technology has matured to meet the demanding industrial requirements in three-dimensional (3D) sensing and LiDAR[34], high-speed optical communications[35] and laser printing[36]. Our ONN system utilizes (1) micrometre-scale VCSEL transceivers for high-speed (gigahertz) data transmission with phase coherence over the entire array via injection locking, (2) coherent detection for low-energy weighted accumulation and (3) holographic data movement as optical dendritic fanout for parallel computing. We experimentally achieve the best-in-class ONN with (C1) full-system energy efficiency (including the energy of optics and the proposed digital electronics) reaching 7 fJ/OP, (C2) compute density potentially exceeding 6 teraOP mm^{-2} s^{-1}) and (C3) inline nonlinearity based on homodyne detection with instantaneous response. Furthermore, the system is (C4) scalable through existing mature wafer-scale fabrication processes and photonic integration, while high-speed (>gigahertz) time multiplexing enables the system to (C5) freely scale to run models with up to tens of billions of neurons (a model with 79,400 parameters is demonstrated, which is 100 times larger than other integrated ONNs, Supplementary Table III).

Results

Schematic

Our VCSEL-ONN architecture consists of a sequence of *N* layers (Fig. 1b). Each layer computes a matrix-vector multiplication $X^n_{(1\times i)} W^n_{(i\times j)} = Y^n_{(1\times j)}$ followed by a nonlinear activation function $f_{\mathrm{NL}}(\cdot)$. Our scheme imitates the 'axon-synapse-dendrite' architecture in biological neurons. As shown in Fig. 1c, we encode the input vector $X^n_{(1\times i)}$ in *i* time steps to the amplitude or phase of a coherent laser oscillator (labelled 'axon'), whose beam is dendritically fanned out to *j* copies for parallel processing. We map the weight matrix $W^n_{(i\times j)}$ in *i* time steps

Table 1 | Figures of merit of VCSEL-ONN

Criteria	Description	VCSEL-ONN
C1	Energy efficiency, ϵ	7 fJ/OP
C2	Compute density, ρ	6 teraOP mm^{-2} s^{-1}
C3	Inline nonlinearity	Instantaneous response
C4	Hardware scalability	Wafer-scale volume production
C5	Model size	79,400 parameters

with phase encoding $\sin[\phi_{W_j}(t)] \propto W_{ij}$ using a 'weight server' consisting of j laser transmitters. Each weighting laser beats with a copy of the input laser on a photo-receiver, producing the homodyne product between the two laser fields (Fig. 1d), as detailed in Supplementary Section I. The resulting photocurrent is accumulated over i time steps, yielding

$$I_j \propto \sum_i A_{W,ij} A_{X,i} \sin(\phi_{W,ij} - \phi_{X,i}) \quad (1)$$

where $A_{X,i}$ and $\phi_{X,i}$, $A_{W,ij}$ and $\phi_{W,ij}$, respectively, are the amplitude and phase of the input and weight lasers. Homodyne multiplication allows linear multiplication $f_L(\cdot) \propto A_X(t)\sin[\phi_W(t)] = X_i W_{ij}$ when the input data is amplitude-encoded $A_X(t) \propto X_i^n$, or the nonlinear operation $f_{NL}(\cdot) \propto \sin[\phi_W(t) - \phi_X(t)] = W_{ij}\sqrt{1-X_i^2} - X_i\sqrt{1-W_{ij}^2}$ with phase-encoding $\sin[\phi_X(t)] \propto X_i^n$(Methods). The input–output response of the two data modulation schemes is modelled as in Supplementary Fig. I.

With the phase-encoding scheme, we incorporate the detection-based optical nonlinearity in our VCSEL-ONN. As shown in Supplementary Fig. Ib, programming the phase of the weight laser tunes the strength of our homodyne nonlinearity. The effectiveness of our homodyne nonlinearity is verified in neural network training, showing a performance similar to that of the rectified linear unit (ReLU) nonlinear activation in handwritten digit and letter classification, as well as fashion product classification (Supplementary Table I). As homodyne detection relies on the photoelectric effect, where an electron is elevated to the conduction band by the absorbed photon, the process is nearly instantaneous, with a time delay of tens of attoseconds[37]. The resulting latency is as short as the optical pulse per symbol, which can be below a femtosecond in principle. This is in contrast to the nanosecond delay with digital[20,27] and electro-optic[21,38] nonlinearities, and cavity- or atom-based optical nonlinearities[23,31]. Its implementation with a photodetector is ultracompact, without instrumental complexity (for example, ultrashort laser pulses[32]).

Based on space–time multiplexing and fanout data-copying, our system is optimized for computing at high density and energy efficiency. It performs matrix-vector multiplication using i time steps and j coherent receivers. With the axon input laser shared among j channels (j-time parallelism), the number of devices scales linearly with $O(j)$, whereas these requirements in CMOS-based microprocessors[16,17] and integrated ONN circuits[20,25] scale quadratically $O(i \times j)$. Our system is thus substantially simplified, with reduced device counts. As batch operations are required in many machine learning tasks, the beams in the weight server can be broadcast for processing a batch of k input vectors simultaneously, as in the simplified schematic shown in Supplementary Fig. 7. Matrix–matrix multiplication $X_{(k\times i)}W_{(i\times j)} = Y_{(k\times j)}$ is enabled with k input encoders, j weight transmitters and i time steps. The device count scales with $O(k + j)$; this would otherwise require $O(i \times j \times k)$ scaling.

Experimental results

Our VCSEL-ONN supports a compact 3D hybrid layout (Fig. 2a) with arrays of VCSELs (A2) bonded on a CMOS driver chip (A1) for data transmission, a phase mask (A3) for beam fanout, and detector arrays (A4) for homodyne multiplication.

To implement the scheme, we engineered a scalable, high-density, phase-stable source with individually addressable VCSEL arrays of tunable, coherent outputs. We fabricated 5 × 5 VCSELs at high density with semiconductor heterostructure microresonators with equal x- and y-direction pitches of 80 μm (Fig. 2d; Methods). All the VCSELs are individually addressable with forward biasing above the lasing threshold using a battery. Each VCSEL emits 100 μW of light with a wall-plug efficiency of 25%. The 3-dB bandwidth of our fabricated devices is ~2 GHz, limited by the photon lifetime at the cavity Q-factor of 10^5 (Supplementary Fig. 9). We exploit VCSELs for analogue data-encoding at a clock rate of 1 giga-symbols per second (GS s^{-1}) with 8 bits of precision: one-billion neurons are activated in 1 s (Fig. 2b). We note that state-of-the art VCSELs with a 3-dB bandwidth of 45 GHz (ref. 39) could further improve our computing rates. The emission wavelength of our VCSELs over the 5 × 5 array is 974.0 ± 0.1 nm. This excellent wavelength homogeneity enables parallel injection locking over the whole array to a leader laser for coherent detection (Methods). An injection optical power of 1 μW per VCSEL is sufficient to achieve a stable phase lock, with a locking range of 1.7 GHz (Fig. 2g). Tuning the individual VCSEL resonance over the locking range, with varying driving voltages, allows phase tuning in the range (−π/2, +π/2), with a π-phase-shift voltage, V_π, of 4 mV (Fig. 2h). Such a low V_π allows phase-only linear modulation with negligible amplitude coupling[40] and negligible crosstalk between neighbouring channels.

We use a single VCSEL array to encode both the input activation and weights. Sharing beam paths improves the interferometric stability in homodyne detection. Among the 25 VCSELs, 24 encode weights forming the weight server, while the corner VCSEL encodes the activations (Fig. 2f). Limited by the large dimension of our diffractive optical element (DOE), the corner laser is separated from the main beams and then fanned out to 9 × 9 spots (Methods). This beam separation is not necessary with pitch-size DOEs, as photonic integration of VCSEL arrays with DOEs for fanout beam copying has been matured for volume production in industry[41]. Each copy of the X^n beam is superimposed to a weight laser beam W_j^n with a beamsplitter. The combined beams are received with a 2D fibre-based detector array, where each detector connects to a switch integrator charge amplifier (Methods) that accumulates the homodyne photon currents.

We characterized the computing accuracy of homodyne interference in our neural network implementation (Fig. 3). We utilized two injection-locked VCSELs to construct a vector–vector multiplication unit based on our physical system's unique nonlinearity $f_{NL}(X_i, W_{ij})$ (Fig. 3c), which differs from conventional multiplication[42] due to the complex-valued nature of the VCSEL network's outputs. We encode two vectors X^n and W_j^n, each with i = 10,000 normally distributed random values, to a VCSEL at a clock rate of R = 1 GS s^{-1} with peak-to-peak voltage of 4 mV. The signals are a.c. coupled to remove the slow thermal drifts (Methods). The experimental time trace agrees well with the calculation in Fig. 3d. The standard deviation of $y - \tilde{y}$ residuals in Fig. 3e,f reveals a computing accuracy of 98% (~6 bits of precision), limited mainly by the phase instability of the set-up and the frequency response of the injection-locked VCSELs. The accuracy can be improved in the future with better photonic integration and VCSELs of higher bandwidth, although the present accuracy is sufficient for a wide range of machine learning tasks[43].

We deployed a neural network inference on our VCSEL-ONN to classify handwritten digits in the Modified National Institute of Standards and Technology (MNIST) database. To this end, we developed a training algorithm with PyTorch using our unique nonlinear weighting function $f_{NL}(W_j^n, X^n)$. Figure 4a shows the trained three-layer model (size 28 × 28 → 100 → 10 → 10). Each image, with 28 × 28 pixels, is flattened and encoded in 784 time steps to an input VCSEL (Fig. 4b) at a driving voltage of 4 mV. The 100 weight vectors in the first hidden layer

https://doi.org/10.1038/s41566-023-01233-w

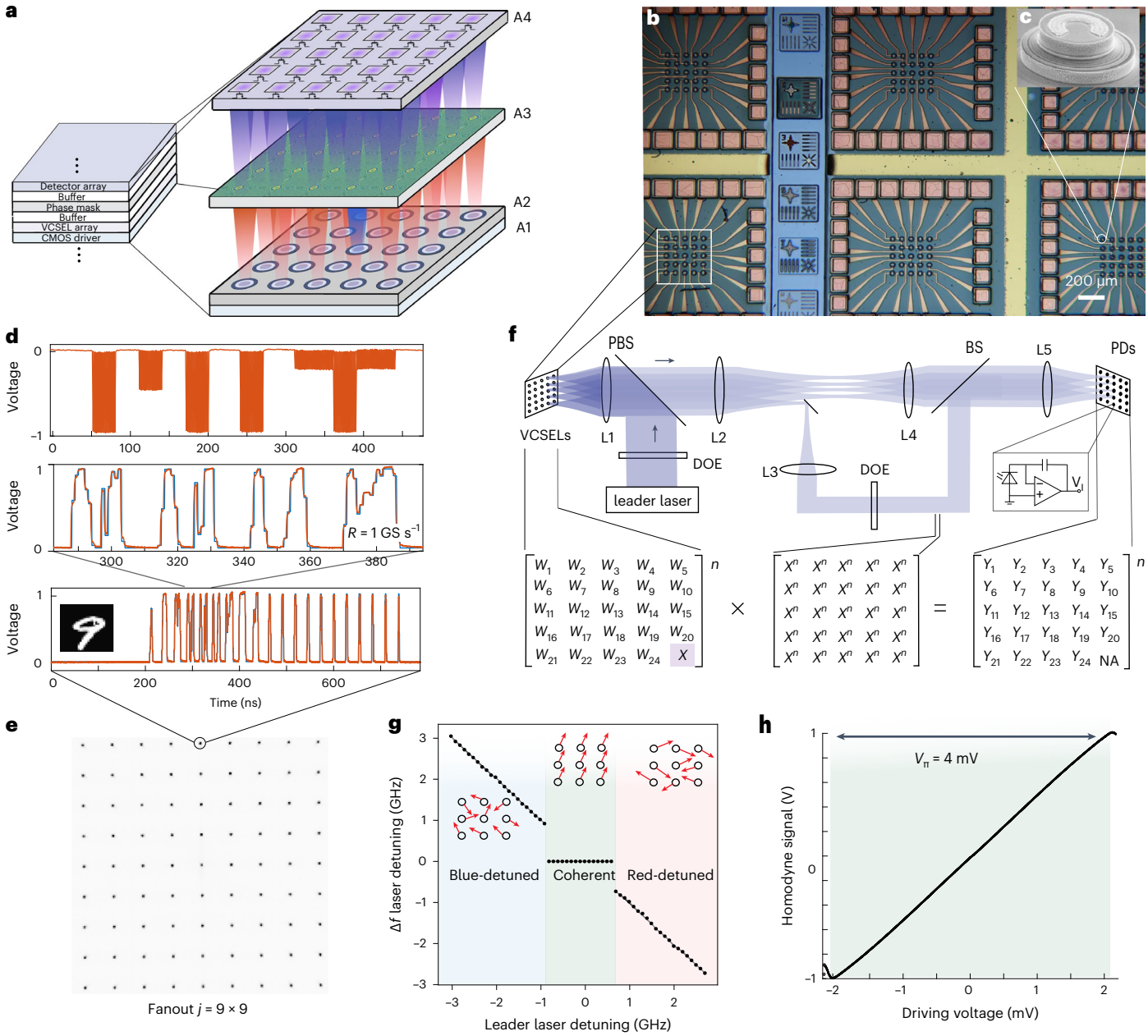

Fig. 2 | Experimental scheme of VCSEL-ONN. a, Proposed architecture with 3D connectivity and photonic integration. In a 2D VCSEL array, the centre (blue) is used as the axon and the others (red) as weight VCSELs. The axon beam is fanned out to j copies, each overlapping a weighted beam onto a photodetector, generating photon currents corresponding to the homodyne product of the two laser fields. **b**, Fabricated VCSEL arrays. Arrays of 5 × 5 wire-bonded VCSELs on a GaAs substrate. **c**, Scanning electron microscopy image of a VCSEL emitter during device processing before adding the top metal connectors. **d**, Analogue data-encoding with an on-chip VCSEL transmitter operating at 1 GS s^{-1}. The MIT logo in the upper plot is constructed with a time sequence of 400 symbols. A 28 × 28-pixel image with a handwritten digit is flattened and encoded in 784 ns. **e**, Optical fanout. A single VCSEL emitting data at 1 GS s^{-1} is fanned out to j = 9 × 9 spots using a diffractive optical element. A camera at the Fourier plane records the resulting beam grid. **f**, Experimental set-up of the VCSEL-ONN apparatus. The input data (X^n, highlighted) is encoded onto the corner VCSELs, and the weight matrix $[W_1^n, \ldots, W_j^n]$ is mapped to the other VCSELs. The input VCSEL is separated from the beam arrays using a beam magnifier (L1 and L2) and D-shaped mirror. DOE, diffractive optical element; BS, beamsplitter; PD, photodetector; PBS, polarizing beamsplitter. **g**, Injection locking range. This is measured by monitoring the beatnote between the leader laser and each VCSEL (Supplementary Section VII). Within the locking range of 1.7 GHz, the VCSELs emit coherently. **h**, Low-V_π and linear modulation. Driving the VCSEL resonance over the locking range allows a phase shift of (−π/2, π/2), V_π = 4 mV.

are flattened and encoded one weight vector per VCSEL. Ideal spatial multiplexing allows processing of 100 weight channels simultaneously; however, limited by the arbitrary waveform generator (AWG) hardware, the data are taken with multiple acquisitions (five VCSELs per acquisition) at a clock rate of 100 MS s^{-1}, although the VCSEL bandwidth allows homodyne interference at >1 GS s^{-1} (Fig. 3c). By switching the AWG channels and translating the VCSEL chips to different arrays in the x and y directions, a total of 100 VCSEL devices from five arrays are used to compute in the first layer. The interference signal between the image data and a weight vector is compared to the digitally calculated result in Fig. 4b. We obtained ~6 bits of compute precision, similar to the result in Fig. 3b. The signal-to-noise ratio (SNR) in the time trace is 135, limited

https://doi.org/10.1038/s41566-023-01233-w

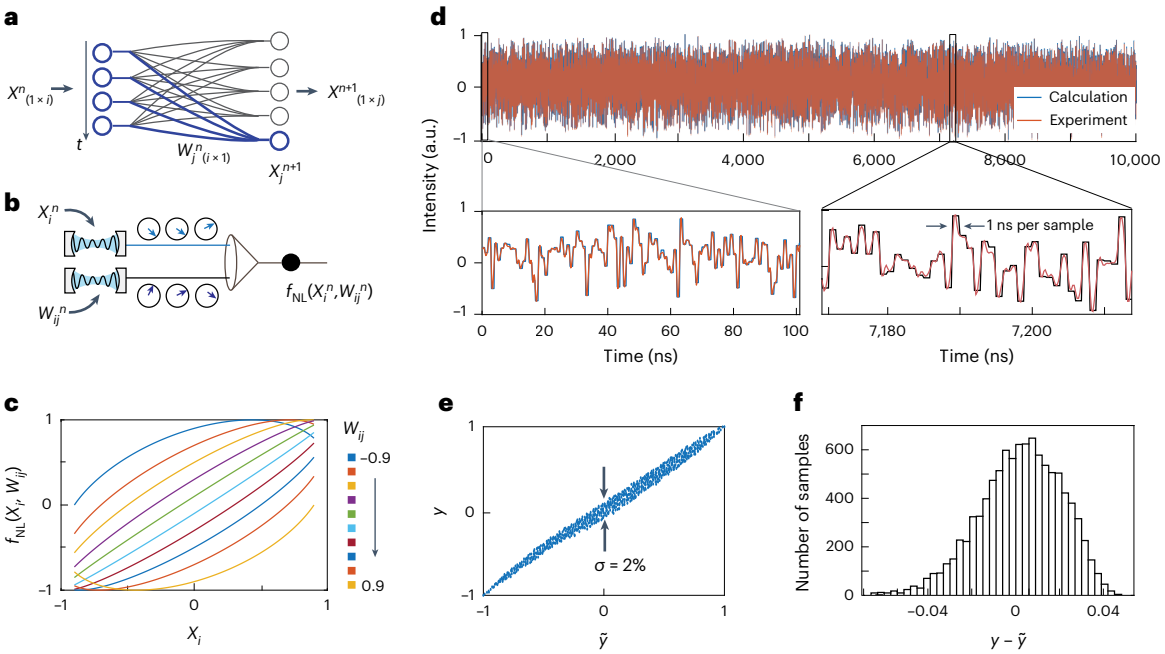

Fig. 3 | Characterization of compute accuracy. a, A neural network nonlinear compute unit, highlighted in blue. It performs vector–vector multiplication $X_j^{n+1} = \sum_i f_{NL}(X_i^n, W_{ij}^n)$ with input vector $X_{1\times i}^n$ and the jth weight vector $W_j^n(i \times 1)$. **b**, Each vector is phase-encoded in time steps to the output of a VCSEL. The interference between two VCSEL fields yields nonlinear weighting $f_{NL}(W_{ij}, X_i) \propto W_{ij}\sqrt{1-X_i^2} - X_i\sqrt{1-W_{ij}^2}$. **c**, Simulation result of homodyne interference with two phase-encoding laser fields results in a nonlinear response. The strength of the nonlinearity increases at higher weights. **d**, Homodyne product of two normally distributed random values at a clock rate of 1 GS s^{-1}. **e**, Floating-point computation shows high accuracy with an error of less than $\sigma = 2\%$. **f**, Histogram of compute errors over 10,000 input samples.

by the photon shot noise (Supplementary Fig. 4). The photocurrent at each channel is accumulated over time with a custom-made time integrator (Methods). The integrated values from the 100 channels are serialized, forming an input vector feeding into the second hidden layer. The weights in the second hidden layer are implemented with ten weighting VCSELs and the interference signal is integrated. Figure 4c shows the real-time integration of the interference signal for processing an image in layer 2. The image classification is read out by the max integrating voltage of the ten VCSEL channels (Fig. 4c,d). Running inference over a dataset of 1,000 MNIST test images, with a total of 158.8 million operations, we obtain an accuracy of (93.1 ± 2.0)%, which is 98% of the model's accuracy in the simulation (95.1%).

System performance

Energy efficiency. Our system enables efficient computing with low-energy VCSEL transmitters and optical parallelism. Supplementary Table II summarizes the energy consumption of each component. The clock rate of 1 GS s^{-1} used in the calculation is demonstrated in Fig. 3. Due to the ultralow $V_\pi = 4$ mV operation, data-encoding with a VCSEL modulator consumes less than 4 nW (corresponding to 4 aJ per symbol at 1 GS s^{-1}), which is six orders of magnitude more efficient than that of previous ONN schemes with thermal phase shifters[20], microring resonators[21], optical attenuators[22] and EO modulators[24,27], which operate with several milliwatts of electrical power. The main optical energy consumption in our system is for laser generation, and we note that VCSEL sources are efficient laser generators with a wall-plug efficiency of 25% in our demonstration and over 57% in record[44]. Our optical energy efficiency, including the electrical power for laser generation and data modulation, is 2.5 fJ/OP (Methods), which is 140× and 1,000× better than the state-of-the-art integrated ONNs of ref. 22 and ref. 24, respectively (Fig. 5).

Further energy costs arise from the electronic digital-to-analogue converters (DACs), analogue-to-digital converters (ADCs), signal amplification and memory access. The energy for the DACs and memory access per use is reduced by a factor of j due to spatial parallel processing with laser fanout. The readout electronics, including ADCs, trans-impedance amplifiers and integrators, is triggered only once after time integration. The energy cost per use is amortized by a total of $2i$ operations in between. Accordingly, the full-system energy efficiency, including both optical consumption and the proposed electronics, can reach 7 fJ/OP (Supplementary Table II), which is more than 100× better than state-of-the-art electronic microprocessors (Fig. 5). Similar to the fanout of the input laser, the weight server can be spatially fanned out (with a factor of k) for batch operations, which reduces the energy for weighing by k folds (Supplementary Section V).

Potential compute density. High compute density is achievable based on compact and dense VCSEL arrays in the 3D architecture. VCSELs are excellent candidates for high-density computing, with a pitch of 80 μm per fabricated device. With optimized optoelectronic integration and pitch matching, the compute density in our system could reach $\rho = 6$ teraOP mm^{-2} s^{-1} (Methods), which is ~20× higher than that of its electronic counterparts (Fig. 5), where improving the throughput density is fundamentally challenging due to limited heat dissipation per chip area. In other ONN configurations, high throughput density requires tiling of photonic devices[20] at high density, which often leads to crosstalk between neighbouring channels and decreased compute accuracy. The channel crosstalk in our VCSEL-ONN is eliminated with VCSEL modulators with ultralow V_π.

Latency. Ultralow latency for nonlinear activation is achieved by incorporating detection-based nonlinearity. In our scheme, each detection event generates photon currents instantaneously, and the photon currents are accumulated in the time integrator for i time steps before being read out. The transit time for photon electrons moving from the photodiode to the charging capacitor, which leads to latency in

https://doi.org/10.1038/s41566-023-01233-w

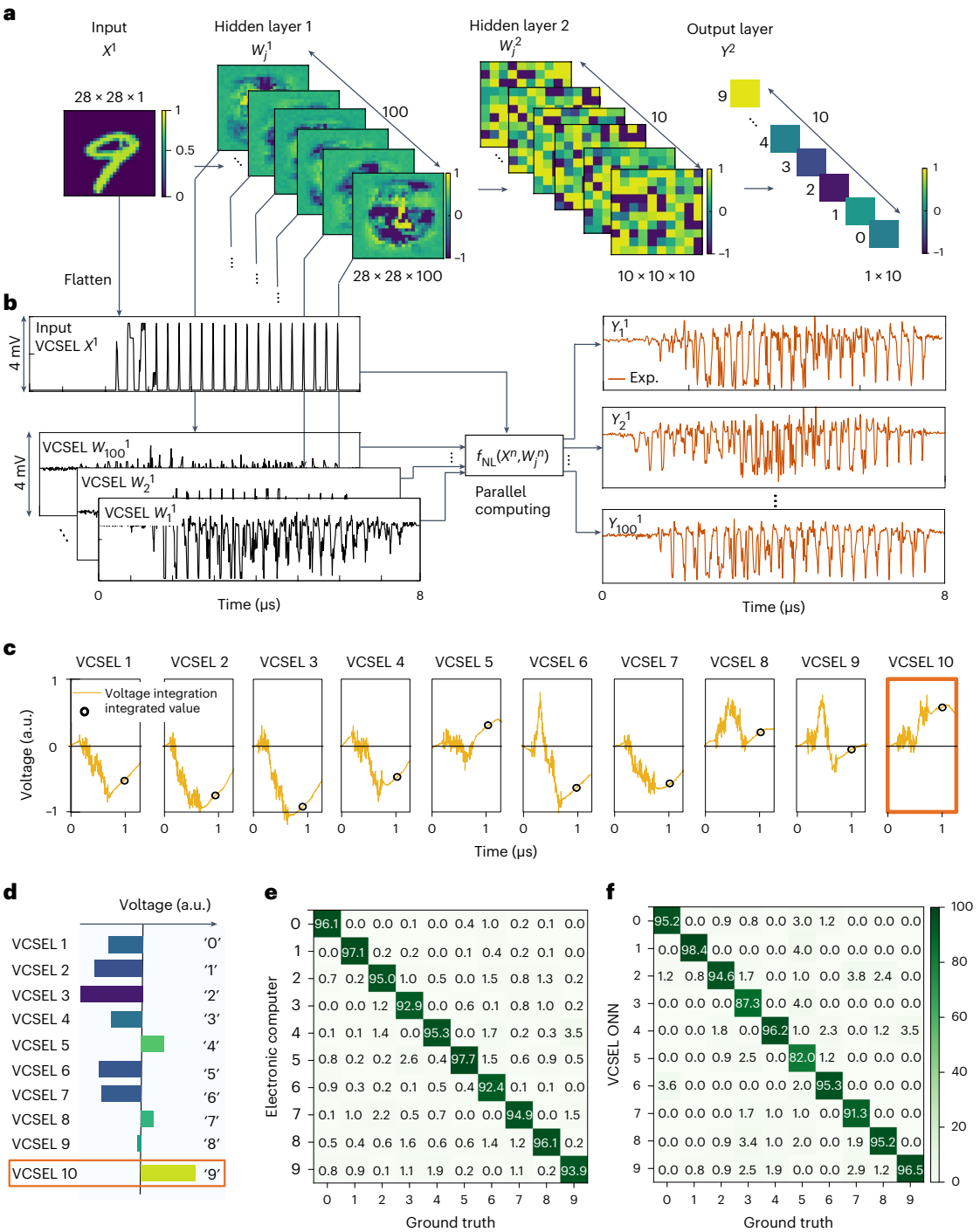

Fig. 4 | Benchmarking of machine learning inference with VCSEL-ONN. **a**, Model for MNIST image classification trained with our unique nonlinearity $f_{NL}(\cdot)$. This consists of one input layer, two hidden layers and an output layer. There are 100 and 10 neurons, respectively, in the first and second hidden layers. **b**, Example parallel multiplication. The input image in layer 1 is flattened and encoded in time steps to the phase of the X^n VCSEL. The weight matrix with 100 vectors is encoded to 100 individual weighting VCSELs. Parallel multiplication results in matrix-vector multiplication (blue) from 100 readout channels. **c**, Example time integration of the interference signal in layer 2. The black dots are the integrated results. **d**, Output layer. The result of MNIST classification is read out by comparing the integrated voltage (black circles in **c**) of the ten processing VCSEL channels. **e**,**f**, Comparison of electrical (**e**) and optical (**f**) confusion matrices, showing good agreement between the ONN and the von Neumann electronic computer systems.

standard photodetectors, is negligible compared to the integration time. So, the latency due to nonlinear activation is negligible. The processing time is dominated by the data-encoding and time integration, which could be as short as 30 ns for a full-size MNIST image at a clock rate of $R = 25\ \mathrm{GS\ s^{-1}}$ (Supplementary Fig. 12).

Discussion

By harnessing the powerful scalability of the VCSEL platform, we have demonstrated a homodyne-based ONN using more than 100 coherent VCSEL transmitters. Our benchmarking of digit classification achieved an accuracy of 93.1% (over 98% of ground truth). With optimized

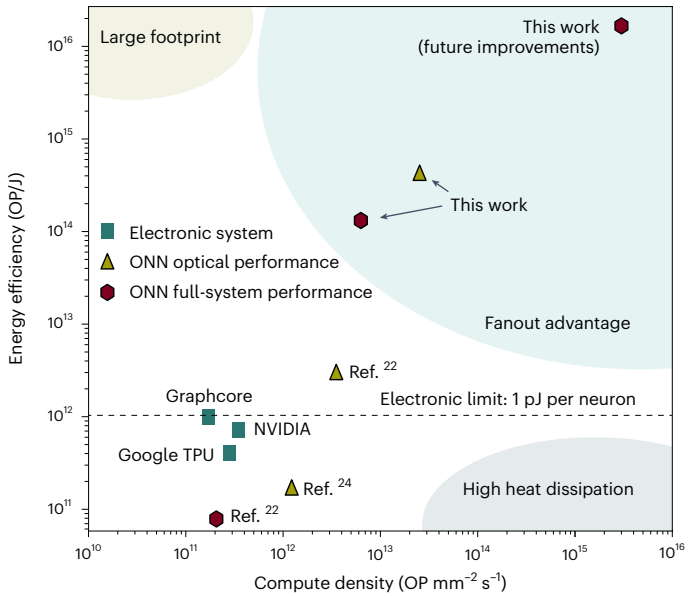

Fig. 5 | Comparison of state-of-the-art neural network accelerators. The electronic systems Google TPU, NVDIA GPU and Graphcore are ASICs optimized for deep learning tasks, with energy efficiency and compute density reaching 1 pJ/OP (Graphcore IPU Gen-2) and 0.35 teraOP mm^{-2} s^{-1} (NVDIA A100). For the ONN optical performance, the energy efficiency accounts for the electrical power in laser generation and data-encoding, and the compute density is calculated from the chip area for matrix operations. For ONN full-system performance, the energy consumption and compute density account for laser generation, data-encoding, nonlinear activation, data readout, signal amplification, ADCs, DACs and memory access. In this work, the energy bound due to electronics at ~1 pJ per neuron is reduced with spatial fanout and time-domain fan-in (Supplementary Table II). The compute densities of this work are potential values, estimated with present-day technology of photonic integration and electronic packaging. The performances of the ONN techniques are summarized in ref. 22 and plotted here. The data in the plot are listed in Supplementary Table IV.

electronics and photonic packaging, the full-system energy efficiency and compute density could reach 7 fJ/OP and 6 teraOP mm^{-2} s^{-1}, respectively, which are improvements of 100× and 20× compared to digital hardware. In the near term, the weight server could be broadcast to k copies for matrix–matrix multiplication $X_{(k\times i)}W_{(i\times j)} = Y_{(k\times j)}$, where the overall throughput scales as $2\times j\times k\times R$, which could exceed 50 petaOP s^{-1} with practical hardware parameters ($j = k = 1{,}000$ and $R = 25$ GS s^{-1}; Supplementary Fig. 12). Such a system is expected to run multiply-accumulate operations with an efficiency of ~50 aJ/OP, limited by the memory access rather than optical energy consumption (Supplementary Table II), and a full-system compute density of 2 petaOP mm^{-2} s^{-1} is within reach due to the high clock rate and large fanout factor (Methods). The lower bound of our optical energy consumption is set by the number of photons required to produce accurately weighted accumulation. Operating large neural network models (with $i > 10^6$), the system may allow less than 1 photon per OP with a standard photodetector at the room-temperature thermal-noise limit (Supplementary Fig. 4).

Our architecture allows compact system integration. The proposed 3D scheme (Fig. 2) can be monolithic and robust against vibrations and environmental noise when the buffer layers for beam propagation are filled with optically transparent materials. The thickness of the buffer layers can be small, on the millimetre scale at large fanout factors ($j = 32\times 32$; Methods). The system also supports in-plane integration with VCSELs bonded to optical waveguides[45,46], with on-chip beamsplitters for spatial fanout and integrated detectors for homodyne multiplication. Our VCSEL-ONN with ultralow-voltage operation is compatible with CMOS digital electronics for optoelectronic interfacing[47,48]. Due to the ultralow energy consumption of the VCSELs (<5 aJ per symbol), rapid programmability with high-speed updating of weights in our system is readily applicable to real-time neural network training.

Although the exponential scaling of neural network models has outpaced the development of electronic processors, this new type of optoelectronic processor with orders-of-magnitude improvement may enable us to continue the scaling of computing power in the post-Moore's law age.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41566-023-01233-w.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (eds Pereira, F., Burges, C., Bottou, L. & Weinberger, K.) Vol. 25 (Curran Associates, 2012); https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
3. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. & Poggio, T. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 411–426 (2007).
4. Young, T., Hazarika, D., Poria, S. & Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **13**, 55–75 (2018).
5. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
6. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
7. Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **71**, 361–390 (2020).
8. Xu, X. et al. Scaling for edge inference of deep neural networks. *Nat. Electron.* **1**, 216–222 (2018).
9. Brown, T. B. et al. Language models are few-shot learners. In *Proc. of the 34th International Conference on Neural Information Processing Systems (NeurIPS)* 1877–1901 (2020); https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
10. Dennard, R. et al. Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE J. Solid-State Circuits* **9**, 256–268 (1974).
11. Horowitz, M. 1.1 Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* 10–14 (IEEE, 2014).
12. Nahmias, M. A. et al. Photonic multiply-accumulate operations for neural networks. *IEEE J. Select. Topics Quantum Electron.* **26**, 1–18 (2020).
13. Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
14. Zhou, H. et al. Photonic matrix multiplication lights up photonic accelerator and beyond. *Light Sci. Appl.* **11**, 30 (2022).
15. Keckler, S. W., Dally, W. J., Khailany, B., Garland, M. & Glasco, D. GPUs and the future of parallel computing. *IEEE Micro* **31**, 7–17 (2011).
16. Jouppi, N. P. et al. In-datacenter performance analysis of a tensor processing unit. In *Proc. 44th Annual International Symposium on Computer Architecture* 1–12 (Association for Computing Machinery, 2017); https://doi.org/10.1145/3079856.3080246

17. Chen, T. et al. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning. *SIGARCH Comput. Archit. News* **42**, 269–284 (2014).
18. Sze, V., Chen, Y.-H., Yang, T.-J. & Emer, J. S. Efficient processing of deep neural networks: a tutorial and survey. *Proc. IEEE* **105**, 2295–2329 (2017).
19. Miller, D. A. B. Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
20. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
21. Tait, A. N. et al. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **7**, 7430 (2017).
22. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
23. Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H. & Pernice, W. H. P. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
24. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
25. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
26. Zhou, T. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photon.* **15**, 367–373 (2021).
27. Xu, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
28. Sludds, A. et al. Delocalized photonic deep learning on the internet's edge. *Science* **378**, 270–276 (2022).
29. Wang, T. et al. An optical neural network using less than 1 photon per multiplication. *Nat. Commun.* **13**, 123 (2022).
30. Wang, C. et al. Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages. *Nature* **562**, 101–104 (2018).
31. Zuo, Y. et al. All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132–1137 (2019).
32. Li, G. H. et al. All-optical ultrafast ReLU function for energy-efficient nanophotonic deep learning. *Nanophotonics* **12**, 847–855 (2022).
33. Tait, A. N. et al. Silicon photonic modulator neuron. *Phys. Rev. Appl.* **11**, 064043 (2019).
34. Kim, I. et al. Nanophotonics for light detection and ranging technology. *Nat. Nanotechnol.* **16**, 508–524 (2021).
35. Liu, A., Wolf, P., Lott, J. A. & Bimberg, D. Vertical-cavity surface-emitting lasers for data communication and sensing. *Photon. Res.* **7**, 121–136 (2019).
36. Koyama, F. Recent advances of VCSEL photonics. *J. Lightwave Technol.* **24**, 4502–4513 (2006).
37. Ossiander, M. et al. Absolute timing of the photoelectric effect. *Nature* **561**, 374–377 (2018).
38. Tait, A. N. et al. Silicon photonic modulator neuron. *Phys. Rev. Appl.* **11**, 064043 (2019).
39. Heidari, E., Dalir, H., Ahmed, M., Sorger, V. J. & Chen, R. T. Hexagonal transverse-coupled-cavity VCSEL redefining the high-speed lasers. *Nanophotonics* **9**, 4743–4748 (2020).
40. Hoghooghi, N., Ozdur, I., Akbulut, M., Davila-Rodriguez, J. & Delfyett, P. J. Resonant cavity linear interferometric intensity modulator. *Opt. Lett.* **35**, 1218–1220 (2010).
41. *Using VCSELs in 3D Sensing Applications* (Finisar Corporation, 2022); https://www.semiconchina.org/Semicon_China_Manager/upload/kindeditor/file/20190415/20190415103954_498.pdf
42. Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M. & Englund, D. Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2019).
43. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R. & Bengio, Y. Quantized neural networks: training neural networks with low precision weights and activations. *J. Mach. Learn. Res.* **18**, 6869–6898 (2017).
44. Jager, R. et al. 57% wallplug efficiency oxide-confined 850 nm wavelength GaAs VCSELs. *Electron. Lett.* **33**, 330–331 (1997).
45. Kumari, S. et al. Vertical-cavity silicon-integrated laser with in-plane waveguide emission at 850 nm. *Laser Photon. Rev.* **12**, 1700206 (2018).
46. Yang, Y., Djogo, G., Haque, M., Herman, P. R. & Poon, J. K. S. Integration of an O-band VCSEL on silicon photonics with polarization maintenance and waveguide coupling. *Opt. Express* **25**, 5758–5771 (2017).
47. Atabaki, A. H. et al. Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip. *Nature* **556**, 349–354 (2018).
48. Sun, C. et al. Single-chip microprocessor that communicates directly using light. *Nature* **528**, 534–538 (2015).

Methods

Device fabrication

The VCSEL arrays were fabricated with a small pitch of 80 μm to maximize the device density. The VCSEL cavities are based on semiconductor heterostructure microresonators with two AlGaAs/GaAs distributed Bragg reflectors (DBRs) as cavity mirrors and a stack of InGaAs quantum wells as the gain medium[49]. The 5 × 5 cavity arrays were patterned by UV lithography and etched by an inductively coupled plasma reactive ion beam. Each cavity, with an outer diameter of 30 μm, was oxidized to an aperture of 4.5 μm to suppress higher-order transverse modes. To improve the laser stability, the entire chip was clad with a polymer layer, and the areas of the VCSEL cavities were reopened. The Au-deposited p-contact of each VCSEL was connected to a signal pad, which was wire-bonded to a printed circuit board linked to external drivers. All the VCSELs share a common ground (golden bars in Fig. 2b). The cross-section of the VCSELs was designed with 1% ellipticity, which allows a polarized laser output with an improved extinction ratio.

Homodyne multiplication

Homodyne interference allows linear and nonlinear operations based on equation (1). The lasers in the weight server are phase-modulated with $\sin[\phi_W(t)] \propto W_{ij}$. Linear operation is activated when the input vector is amplitude-encoded with $A_X(t) = X_i$ (Supplementary Fig. Ia). A_W is constant because the VCSELs for encoding weights are phase-only modulators. When the input and weight lasers are biased in phase ($\phi_X = 0$), the interference signal is simplified to $\Delta I(t) \propto A_X(t)\sin[\phi_W(t)] = X_i W_{ij}$. Nonlinear operation is allowed when the input vector is phase-encoded with $\sin[\phi_X(t)] \propto X_i$. The generated photocurrent on the detector is $\Delta I(t) \propto \sin[\phi_W(t) - \phi_X(t)] = W_{ij}\sqrt{1-X_i^2} - X_i\sqrt{1-W_{ij}^2}$, where mappings $\phi_X(t) = \sin^{-1}(X_i)$ and $\phi_W(t) = \sin^{-1}(W_{ij})$ are used. The input–output response of the linear and nonlinear models is simulated in Supplementary Fig. Ib. In linear operation, when both laser fields are phase-only modulated, the non-interference terms are direct currents that can be decoupled from the a.c. terms, so the inference can be detected with an unbalanced single detector for system simplicity.

Injection locking

As shown in Fig. 2f, the beam of the leader laser, passing through a DOE, is reflected to the VCSEL arrays using a polarizing beamsplitter (PBS). At the Fourier plane of the coupling lens, the leader laser splits to a beam grid of spacing equal to the pitch of the VCSELs. The polarization of the PBS is aligned at 45° with respect to that of the VCSELs. So, half of the leader laser power is coupled to the VCSEL cavity, locking the phase of the remaining VCSEL oscillators. The other half, being reflected by the VCSEL front DBR, is rejected by the PBS to avoid falling onto the homodyne detectors and leading to undesired interference. Such an injection-locking technique has been demonstrated recently to achieve a high rejection ratio[50]. To simultaneously injection-lock the whole array, we tuned all the VCSELs to the wavelength of the leader laser by varying the bias voltage using a battery-supplied d.c. controller on each VCSEL channel. The injection lock was confirmed by monitoring the beatnote between the leader laser and each individual VCSEL. More details are provided in Supplementary Section VII.

Phase modulation with injection-locked VCSELs

The phase (ϕ) of an injection-locked VCSEL is given by the frequency detuning between the leader laser and the VCSEL's free-running frequency, $\sin(\phi) \propto \delta_d/\delta_r$, where δ_d is the detuning and δ_r is the injection-locking range. Consistent with the theory of injection locking, the range δ_r is proportional to the square root of the injecting power (Supplementary Fig. 10), so a small V_π (millivolt range) is achieved by reducing the injecting power (to ~1 μW per VCSEL). The result is similar to that in ref. 40. The frequency response of the injection-locked VCSELs in the thermal region (<10 MHz) is ~10-dB stronger than that in the free-carrier region (Supplementary Fig. 10d). To decouple from the thermal effect, we modulated the data with a high-frequency local oscillator and demodulated it from the homodyne signal (Supplementary Fig. 11). This data modulation scheme is not needed with VCSELs operating at higher data rates, as experimentally demonstrated in ref. 51.

Experimental set-up

In Fig. 2f, the 5 × 5 VCSEL (pitch 80 μm) emission beams are shown coupled out of chip using a beam expander (f_{L1} = 40 mm and f_{L2} = 400 mm). At the focus of L2, a D-shaped mirror separates the input X^n beam from the weight beams. The X^n beam is collimated with a lens f_{L3} = 400 mm and fanned out using a DOE (MS-259-K-Y-X, Holo/OR Ltd) with diffraction angle $\theta = 0.11(1)°$. At the Fourier plane of L5 (f_{L5} = 75 mm), the X^n beam splits to 9 × 9 beam arrays, and the centre 5 × 5 spots are coupled to a 2D fibre array with a pitch of 150 μm (Beijing Reful Co. Ltd). The weight beams were individually coupled to the same fibre array using a beam compressor formed by f_{L4} = 400 mm and f_{L5} = 75 mm.

Time-integrating receiver

A time-integrating charge amplifier is used to accumulate homodyne photon currents. This device was homemade and is based on an integrated circuit (IVC102, Texas Instruments). Photon currents generated from the detector are deposited on the capacitor. The voltage across the capacitor is $V_I = \int \mathrm{d}t \frac{I(t)}{C} \propto \sum_i f(X_i, W_{ij})$, where C = 10 pF is the capacitance used here. The voltage is digitized after integrating over i time steps and the capacitor is discharged after each readout. The time-integrating receiver allows an i-fold lower readout speed than the multiplication rate. This substantially reduces the bandwidth requirements on the receivers. For example, with $i = 10^6$, an integrating receiver at 1 MHz can potentially read out data multiplying at 1 THz.

Training of machine learning models

The training model is custom-designed to implement tailored nonlinearity. It consists of one input layer, two fully connected hidden layers and an output layer (Fig. 4). The input layer consists of 784 neurons, corresponding to a full-size MNIST image with a handwritten digit. Two fully connected hidden layers are used. In each layer, the matrix-vector multiplication is computed with our custom nonlinear synaptic weighting function $f_{NL}(W_j^n, X^n)$. The output layer consists of ten neurons, and each neuron represents a digit (from 0 to 9). The prediction result of each digit is given by the number of neurons with the largest value. We implement batch normalization[52] in each layer to accelerate the training process. The initial weights[53] are optimized such that the integrated partial sums in each layer converge to a mean value around 0, which allows one to maximize the experimental signal level without being bound by the dynamic range limit. We utilize the cross-entropy loss function and retrieve the gradients in each iteration in a model implemented in PyTorch[54]. A large learning rate is set to start the training and is gradually reduced to optimize accuracy and avoid overfitting. The training model was modified to implement training on handwritten letter classification and fashion product classification, to verify the effectiveness of our nonlinearity on different tasks.

Energy efficiency

The VCSEL emits P_i = 100 μW of optical power while consuming P_b = 400 μW of electrical power. The injection-lock power is P_{INJ} = 1 μW per VCSEL. The power for data modulation is $P_M = V_\pi^2/R_{VCSEL}$ (3.6 nW), with R_{VCSEL} = 4.3 kΩ and V_π = 4 mV. The maximum clock rate of the system is R = 1 GS s^{-1}. With a fanout factor of $j = 9 \times 9$, the optical energy efficiency is $(P_b + P_{INJ} + P_M)/(2jR)$ = 2.5 fJ/OP, which is dominated by the laser power. The full-system energy efficiency can reach 7 fJ/OP with the proposed electronic components discussed in Supplementary Table IV.

Potential compute density

The chip area is dominated by the VCSEL transmitters (A2, 80 × 80 μm^2 per device) because (1) the silicon detector arrays (A4) are compact (for example, detector pixels are commercially available with a pitch size of <0.8 × 0.8 μm^2 for an image sensor), (2) the integration of phase masks and microlenses (A3) on the VCSEL output facet is a mature technology in the industry[41], and (3) an area of 80 × 80 μm^2 can store 38,000 8-bit digital values (A1) (a useful static random access memory (SRAM) cell in 5-nm bulk CMOS at Taiwan Semiconductor Manufacturing Company (TSMC) has dimensions of 0.021 μm^2; ref. 55), and the same footprint also supports more than one million transistors for digital logic, VCSEL driving, DACs, ADCs and so on. The potential optical compute density is $\rho = 2jR/a = 25$ teraOP mm^{-2} s^{-1}, accounting for the input laser being fanned out. For full-system implementation, we assume that the CMOS driver chip (A1), the VCSEL chip (A2), the phase mask chip (A3) and the detector chip (A4), as shown in Fig. 2, are all pitch-matched, resulting in a fourfold increase in chip area and thus a fourfold lower system compute density $\rho_{sys} = 2jR/(4a) = 6$ teraOP mm^{-2} s^{-1}. With future improvements, including upgrading the clock rate ($R = 25$ GS s^{-1}) and fanout factor ($j = 1{,}000$) (Supplementary Fig. 12), the compute density is expected to reach $\rho = 2jR/(4a) = 2$ petaOP mm^{-2} s^{-1}.The chip area efficiency of the weight server can be improved with weight broadcasting (Supplementary Fig. 8).

Potential system compactness

The 3D system can be compact in volume. The vertical spacing between the VCSEL chip and the detectors is mainly limited by the focal length (f) of the lenses used for (1) focusing each VCSEL output to the corresponding detector pixel and (2) Fourier-transforming the phase-masked beam profile (fanout copying). Focusing the VCSEL output (mode field diameter of ~5 μm) can be implemented with a high-numerical-aperture (NA ≈ 1) metalens at the VCSEL output facet ($f < 10$ μm) (ref. 56). For Fourier-transforming the phase-masked beam profile, each VCSEL beam is fanned out to j copies with a spacing to match the pitch ($d = 80$ μm) of the VCSELs (Fig. 2a). The diffraction angle of the mth-order beam from the DOE is $m \times \theta$, with $\tan(m\theta) = m \times d/f$. For the fanout factor $j = 32 \times 32$, the DOE diffracts beams of up to $m = \pm 16$ orders. So a focal length $f = md/\tan(m\theta) = 4.5$ mm is required with diffraction angle $\theta = 1°$ (diffraction angles θ ranging from 0.1 to 10° are commercially available; Holo/OR). The vertical spacing (on the millimetre scale) is similar to the chip size of 32 × 32 VCSELs. Note that the technology for the photonic integration of VCSELs chips, microlens arrays and DOEs is mature for volume production[41].

Data availability

All the data that support the findings of this study are included in the main text and Supplementary Information. The data are available from the corresponding authors upon reasonable request.

References

49. Heuser, T. et al. Developing a photonic hardware platform for brain-inspired computing based on 5×5 VCSEL arrays. *J. Phys. Photon.* **2**, 044002 (2020).
50. Rowland, J., Perrella, C., Light, P., Sparkes, B. M. & Luiten, A. N. Using an injection-locked VCSEL to produce Fourier-transform-limited optical pulses. *Opt. Lett.* **46**, 412–415 (2021).
51. Bhooplapur, S., Hoghooghi, N. & Delfyett, P. J. Pulse shapes reconfigured on a pulse-to-pulse time scale by using an array of injection-locked VCSELs. *Opt. Lett.* **36**, 1887–1889 (2011).
52. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proc. 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research* Vol. 37 (eds Bach, F. & Blei, D.) 448–456 (PMLR, 2015); https://proceedings.mlr.press/v37/ioffe15.html
53. Mishkin, D. & Matas, J. All you need is a good init. In *International Conference on Learning Representations (ICLR)* (2016); http://cmp.felk.cvut.cz/~mishkdmy/papers/mishkin-iclr2016.pdf
54. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) Vol. 32 (Curran Associates, 2019); https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
55. Yeap, G. et al. 5 nm CMOS production technology platform featuring full-fledged EUV, and high mobility channel FinFETs with densest 0.021 μm^2 SRAM cells for mobile SoC and high performance computing applications. In *Proc. 2019 IEEE International Electron Devices Meeting (IEDM)* 36.7.1–36.7.4 (IEEE, 2019).
56. Hadibrata, W., Wei, H., Krishnaswamy, S. & Aydin, K. Inverse design and 3D printing of a metalens on an optical fiber tip for direct laser lithography. *Nano Lett.* **21**, 2422–2428 (2021).

Acknowledgements

This work is supported by the Army Research Office under grant no. W911NF17-1-0527, NTT Research under project no. 6942193 and NTT Netcast award 6945207. I.C. acknowledges support from the National Defense Science and Engineering Graduate Fellowship Program and National Science Foundation (NSF) award DMR-1747426. L.B. is supported by the National Science Foundation EAGER programme (CNS-1946976) and the Natural Sciences and Engineering Research Council of Canada (PGSD3-517053-2018). L.A. acknowledges support from the NSF Graduate Research Fellowship (grant no. 1745302). S.R. acknowledges financial support from the Volkswagen Foundation via the project NeuroQNet 2. We thank S. Bandyopadhyay and C. Brabec of MIT and Y. Cui of TUM for comments on the manuscript. We also thank D.A.B. Miller of Stanford University, N. Harris and D. Bunandar of Lightmatter, and S. Lloyd of MIT for informative discussions.

Author contributions

Z.C., R.H. and D.E. conceived the experiments. Z.C. performed the experiment, assisted by A.S., R.D. and I.C. A.S. conducted high-speed measurements on the VCSEL transmitters and developed the integrating electronics. R.D. created the software model for neural network training. I.C. performed electronic packaging on the VCSEL arrays. L.A. assisted with assembling an initial set-up for testing VCSEL samples. A.S. and L.B. assisted with discussions on the experimental data. T.H., N.H., J.A.L. and S.R. designed and fabricated the VCSEL arrays and characterized their performance. R.H. and D.E. provided critical insights regarding the experimental implementation and results analysis. Z.C. wrote the manuscript with contributions from all authors.

Competing interests

Z.C., D.E. and R.H. have filed a patent related to VCSEL ONNs, under application no. 63/341,601. D.E. serves as scientific advisor to and holds equity in Lightmatter Inc. A.S. is a senior photonic architect at Lightmatter Inc. and holds equity. Other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41566-023-01233-w.

Correspondence and requests for materials should be addressed to Zaijun Chen, Ryan Hamerly or Dirk Englund.

Peer review information *Nature Photonics* thanks Xingyuan Xu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.