nature photonics

Article

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

Deep learning with coherent VCSEL neural networks

In the format provided by the authors and unedited

CONTENTS

I. HOMODYNE DETECTION

Matrix operations in our VCSEL-ONN are based on homodyne detection. A compute unit for vector-vector multiplication composed of two VCSEL emitters is shown in Fig. S1. We encode the input vector X^n (of size $1\times i$) and a weight matrix to W^n (of size $i\times 1$) in i time steps to the amplitude or phase of the VCSEL emission. The two lasers are coherent by means of injection-locking using a leader laser emitting at frequency ω. The electric field of the two laser oscillators is

$$E_X(t) = A_X e^{-i\omega t+\phi_X}, \tag{1}$$

$$E_W(t) = A_W e^{-i\omega t+\phi_W}, \tag{2}$$

where A_X and ϕ_X, A_W and ϕ_W, respectively, are the amplitude and phase of the input laser and the weight laser. The beams of the two VCSELs are overlapped using the beamsplitter, which induces a phase delay of $\pi/2$ to the reflected beam. The homodyne receiver detects photocurrents as

$$I^+ \propto \ [E_W e^{-i\pi/2} + E_X] \times [E_W e^{-i\pi/2} + E_X]^*, \tag{3}$$

$$I^- \propto \ [E_W + E_X e^{-i\pi/2}] \times [E_W + E_X e^{-i\pi/2}]^*, \tag{4}$$

Plugging Eq. 1 and Eq. 2 into Eq. 3 and Eq. 4, one obtains

$$I^+ \propto \ |A_X|^2 + |A_W|^2 + 2A_X A_W \sin[\phi_W - \phi_X], \tag{5}$$

$$I^- \propto \ |A_X|^2 + |A_W|^2 - 2A_X A_W \sin[\phi_W - \phi_X], \tag{6}$$

When balance detection is used, it reads the differential currents,

$$\Delta I(t) = I^+(t) - I^-(t) \propto A_X A_W \sin[\phi_W - \phi_X] \tag{7}$$

where the non-interference terms in Eq. 5 and 6 are canceled and the interference term is enhanced by a factor of 2, due to the opposite phase induced by the beam-splitter. The weight VCSELs are phase encoded with $\sin[\phi_W(t)] \propto W_i$.

Linear operation is activated when the input vector is amplitude encoded with $A_X(t) = X_i$, as shown in Fig. S1a. Because the VCSELs for encoding of weights are phase-only modulators, A_W is constant. When the phase of the input laser is set $\phi_X = 0$ (the input VCSEL has the same phase as the injecting laser), the interference signal is simplified to

$$\Delta I(t) \propto A_X(t) \sin[\phi_W(t)] = X_i W_{ij}, \tag{8}$$

where the multiplication of the input value X_i and a weight value W_{ij} is achieved in the homodyne product.

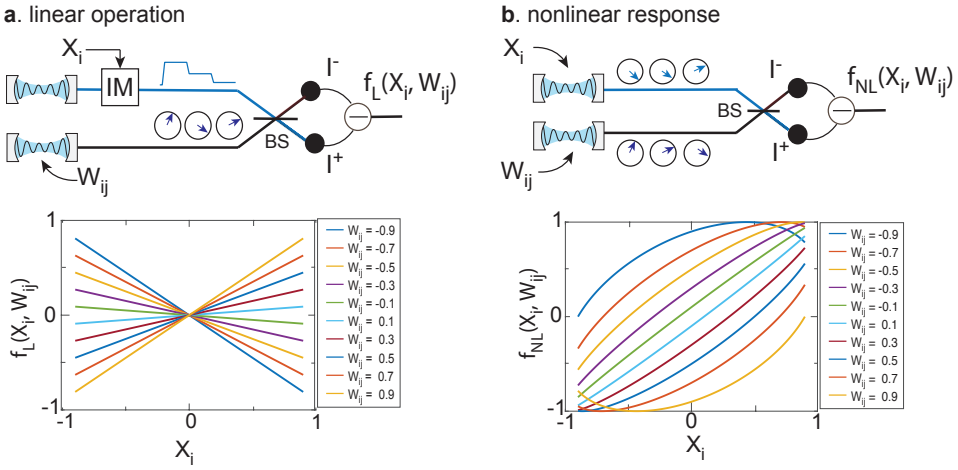
Nonlinear operation is allowed (Fig. S1b) when the input vector is phase encoded, $\sin[\phi_X(t)] \propto X_i$. Here both laser fields are with phase-only modulation, so the non-interference terms are direct currents that can be coupled out from the AC term. The inference can be detected with a balanced detector for signal-to-noise ratio (SNR) improvement, or with a single detector for system simplicity. The generated photocurrent on the detector is

$$\Delta I(t) \propto I(t)^+ \propto \sin[\phi_W(t) - \phi_X(t)] = W_{ij}\sqrt{1 - X_i^2} - X_i\sqrt{1 - W_{ij}^2}, \tag{9}$$

where $\sin[\phi_W(t) - \phi_X(t)] = \sin[\phi_W(t)]\cos[\phi_X(t)] - \cos[\phi_W(t)]\sin[\phi_X(t)]$, $\sin^2(\phi) + \cos^2(\phi) = 1$, $\sin[\phi_X(t)] \propto X_i$ and $\sin[\phi_W(t)] \propto W_{ij}$ are used. The input-output response of the linear and nonlinear models is simulated in Fig. S1.

II. HOMODYNE NONLINEARITY

Homodyne detection allows linear (Eq. 8) and nonlinear operations (Eq. 9), depending on the encoding format of the input data. The response is linear when the input is amplitude modulated and nonlinear when it is phase modulated.

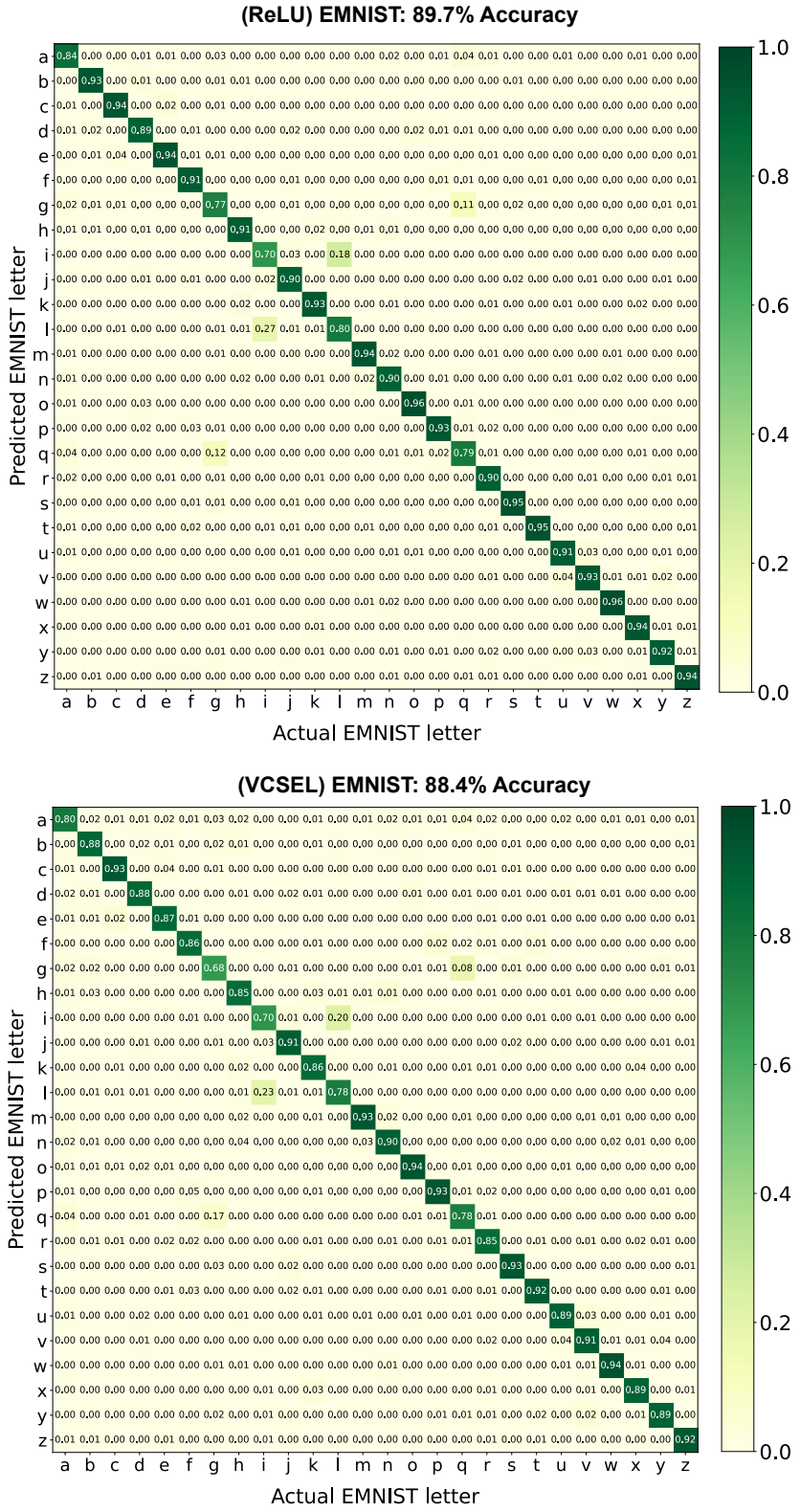


Supplementary Figure 1: Linear and nonlinear operation based on homodyne coherent detection. **a.** Linear input-output response when the input data is amplitude encoded. **b.** Nonlinear activation when the input data is phase encoded. The input-output nonlinearity that increases at high weights is programmable by setting the phase of the weight laser.

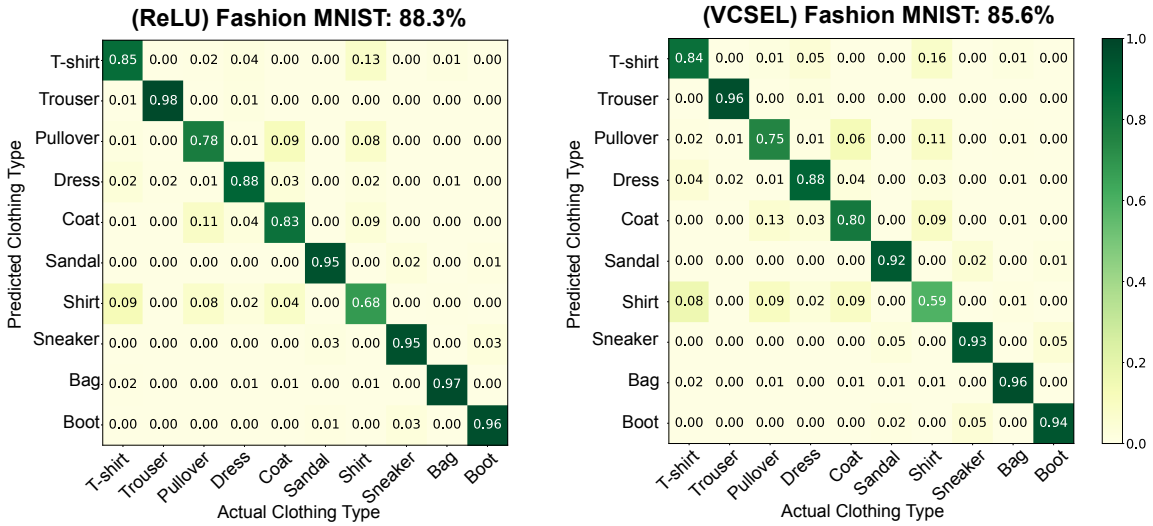
To verify the effectiveness of our nonlinearity, we developed a standard PyTorch training model (Methods), which we applied to image classification on three standard datasets: MNIST (digits 0-9), EMNIST (letters a-z), and Fashion MNIST (10 classes of fashion products). The model was constructed to implement back-propagation training, respectively, with linear matrix multiplication (i.e., no nonlinearity), the ReLU nonlinearity, and the homodyne VCSEL nonlinearity. The models were trained with the same settings: same initial weights, automatically adjusted learning rates, cross-entropy loss function, etc.

Table I: Model accuracy with different activation functions

Dataset	Classes	Model size	Linear	ReLU	Homodyne
MNIST	10	28×28→100→10→10	90.05%	94.13%	95.12%
MNIST	10	28×28→50→100→10	89.85%	96.11%	95.35%
EMNIST	26	28×28→256→128→26	-	87.2%	86.1%
EMNIST	26	28×28→512→256→26	-	89.7%	88.4%
Fashion MNIST	10	28×28→256→128→10	-	88.3%	85.6%



Supplementary Figure 2: training results of hand-written letter classification. 26 classes classification for 26 letters. The model of VCSEL homodyne nonlinearity has similar accuracy as that of a general neural network with linear multiplication + ReLU nonlinearity. The model size of the networks is $28\times28\rightarrow512\rightarrow256\rightarrow26$.



Supplementary Figure 3: Training results of Fashion MNIST classification. 10 Classes are classified. The model size of the network is $28\times28\rightarrow256\rightarrow128\rightarrow10$.

III. ENERGY CONSUMPTION

A fundamental limit to the optical energy consumption is given by the optical power required to achieve the desired signal-to-noise ratio (SNR) in homodyne detection, which sets the compute bits of precision. In this section, we model the signal and noise sources in the system and validate the model with our experimental results. Then we apply the model to discuss the lower bound of optical energy consumption for our current system limit and future improvement.

1. Signal-to-noise analysis

The SNR of optical interference with two laser fields has been well understood in Ref. [1] and is adapted here. We denote P_X and P_W as the power of the input laser and the weight laser on the homodyne receiver. The root mean square amplitude of the interference signal is $S = \sqrt{2}\eta\sqrt{P_X P_W} = \sqrt{2\gamma}\eta P_i$, where η is the quantum efficiency of the photo-detector. For simplicity, we denote $\gamma = P_W/P_X$ and $P_X = P_i$. We include the detector thermal noise, the photon shot noise and the laser intensity noise in the model. The total noise amplitude is

$$N_t = \sqrt{[(\eta \mathrm{NEP})^2 + 2(1+\gamma)h\nu\eta P_i + b(1+\gamma^2)(\eta P_i)^2(\mathrm{RIN})]B} \tag{10}$$

where NEP is the noise equivalent power of the photo-receiver. h is Planck's constant, ν is the frequency, RIN is the relative laser intensity noise. b=1 accounts for the reduction of relative intensity noise due to the balance detection and 2 for unbalanced detection. The noise spectrum is integrated over the effective detection bandwidth B=1/(2T), with T being the acquisition time. The average uncertainty in our homodyne signal is,

$$\sigma_H = N/S_t = \frac{1}{2\sqrt{T}}\left[\frac{\mathrm{NEP}^2}{\gamma P_i^2} + \frac{4c_\gamma h\nu}{\eta P_i} + 2bc_{\gamma^2}(\mathrm{RIN})\right]^{-1/2} \tag{11}$$

where $c_\gamma = (1+\gamma)/(2\gamma)$ and $c_{\gamma^2} = (1+\gamma^2)/(2\gamma)$ accounts for the power ratio of the two lasers. The SNR can be calculated as $1/\sigma_H$.

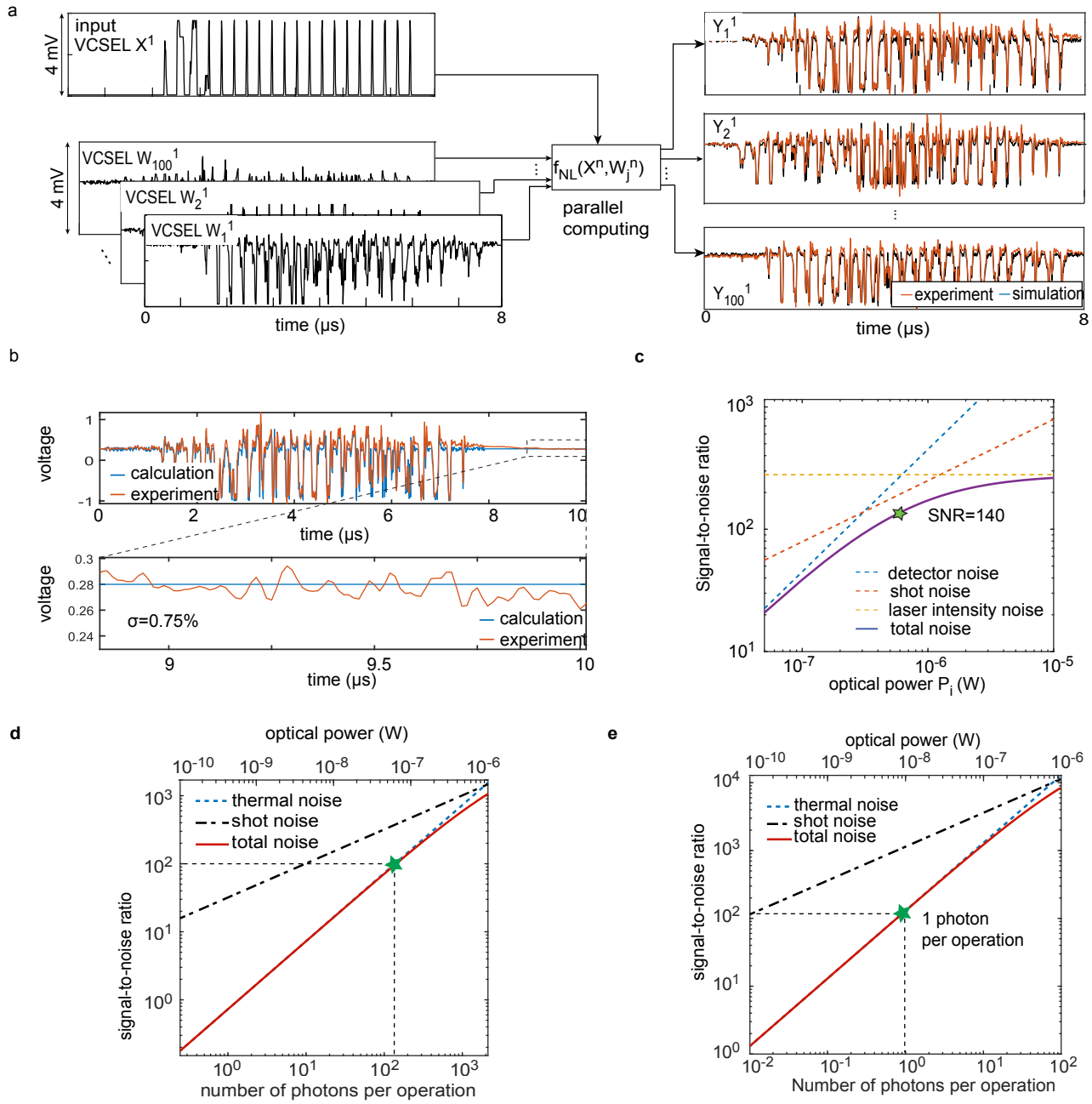
In the implementation of the neural network model (Fig. S4), the data is read out every time step T=t_c=1/R=10 ns. The SNR in the experimental homodyne signal is 135, which agrees well with the theoretical SNR of 140 predicted by the model (Fig. S4). The main contribution of noise source is shot noise.

2. Analysis of system performance

We analyze the system performance utilizing the model above. Different from the readout of every time step in the previous section (Fig. S4), here we use an integrating receiver, which only reads out the integrated value after i time steps. The data acquisition period is $T = i{\cdot}t_c$. The factor of i longer integration time results in a factor of $\sqrt{i}$ improvement in the SNR, as we use amplitude representation. Eq 11 is modified to

$$\sigma_I = N/S = \frac{1}{2\sqrt{i\cdot t_c}}\left[\frac{\mathrm{NEP}^2}{\gamma P_i^2} + \frac{4c_\gamma h\nu}{\eta P_i} + 2bc_{\gamma^2}(\mathrm{RIN})\right]^{-1/2} \tag{12}$$

$$S/N = 2\sqrt{i\cdot t_c}\left[\frac{\mathrm{NEP}^2}{\gamma P_i^2} + \frac{4c_\gamma h\nu}{\eta P_i} + 2bc_{\gamma^2}(\mathrm{RIN})\right]^{1/2} \tag{13}$$



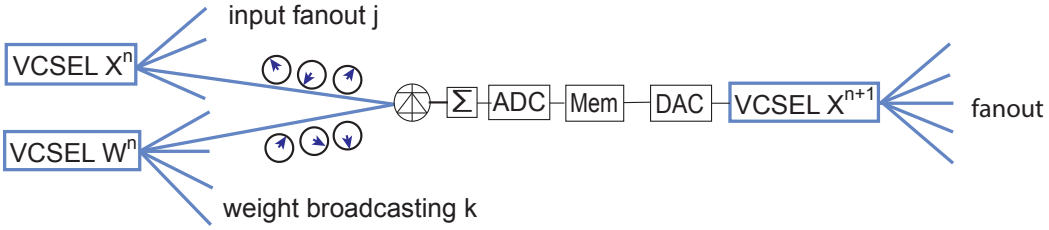
Supplementary Figure 4: SNR analysis. **a.** Experimental demonstration of homodyne interference. A VCSEL encoding an image data X^1 is fanned out to j=9×9 copies. Each copy multiplies a weight vector (encoded in VCSEL W_j). The homodyne signal agrees well with the calculated result. **b.** The SNR in the homodyne signal is given by the standard deviation of the baseline normalized to the peak-to-peak amplitude $1/\sigma$=135. **c.** the SNR is modeled with the experimental conditions: the NEP of the detector is 5 pW/$\sqrt{Hz}$, and the laser relative intensity noise is -145 dBc/Hz, b=2 for unbalanced detection, γ=j=9×9, ν=307.5 THz, and $\eta = 0.65$. The homodyne detector receives P_W=50 μW and P_X=0.6 μ W (50 μW fans out to γ=j=9×9 copies). The data clock rate is R=100 MS/s, corresponding to t_c=1/R=10 ns. **d** and **e** SNR as a function of input laser power (or the number of photons per operation). The SNR is calculated over i integrating time steps. **d** Experimental conditions. The clock rate is R=1 GS/s and the signal is accumulated over i=784 time steps. The detector NEP is 1 pW/$\sqrt{\text{Hz}}$. b=2 for unbalanced detector, γ=1, ν=307.5 THz, and $\eta = 0.65$. **e.** Performance with high-speed VCSELs. With R=25 GS/s and K=10^6, the acquisition time is T=400 ns per readout. Less than 1 photon per operation allows for a compute precision of about 6∼7 bits.

The theoretical lower bound to laser power is given by the number of photons required to produce a homodyne signal with sufficient bits of compute precision, which is ultimately limited by the required SNR from detection. Time integrating receivers (Methods), in contrast to the conventional amplified detector, only read out after accumulating over

i time steps, improving the SNR by a factor of $\sqrt{i}$. With off-the-shelf technology, the thermal noise limit of computing from integration detection is 200 photons/OP (corresponding to 40 aJ/OP).

Fig. S4 **d** and **e** show the evolution of SNR with optical power. In **d**, we analyze the theoretical energy limit with our experimental conditions. The contribution of laser intensity noise in the plotted power range is negligible compared to the detector noise and shot noise. With time integration, the detector bandwidth is reduced by a factor of i (e.g., B=1 MHz is sufficient for R=1 GHz and i=1,000). The thermal noise on the detector is reduced accordingly with $\sqrt{i}$. We use NEP=1 pW/$\sqrt{\text{Hz}}$ in the calculation due to the reduced bandwidth. As our system is designed to operate at room temperature, the SNR at low power level is dominated by thermal noise, which is different from the results from Ref. [2, 3], where the detector is cooled (and that consumes energy) to operate at the shot noise limit. As shown in Fig. S4d, an SNR of 100 is obtained with 200 photons per operation, corresponding to the energy efficiency ϵ=40 aJ/OP. The SNR of 100 corresponds to about 7 bits of precision, which is sufficient for most neural network tasks [4]. In the future, with high-speed VCSELs that operate at R=25 GS/s and with larger model sizes that integrate over a million neurons i=10^6, less than 1 photon per operation (P_i=10 nW) is achievable (Fig. S4e), corresponding to 0.2 aJ/OP.

3. system power consumption



Supplementary Figure 5: Proposed optoelectronic design. A vector-vector multiplication unit.

We analyze the energy consumption of each component in the system, including both the optical energy and the consumption of electronic components, as shown in Table II. A compute unit with the proposed electronic interface is shown as Fig. S5, a copy of the input VCSEL interferes with a copy of the weight VCSEL at the photo-detector, generating photon currents proportional to the homodyne product. The photon currents are accumulated at the integrator. The voltage is digitized with an analog-to-digital converter (ADC) and stored in the memory. The memory is accessed by a digital-to-analog converter (DAC), which drives the input VCSEL at the next layer.

Table II: Energy budget for our experimental apparatus and future improvement using conventional technology.

Parameter	system performance			near-term improvement		
	Value	*Fan-out	†Energy/OP	Value	Fan-out	Energy/OP
VCSEL bias voltage (V_b)	1.3 V					
VCSEL bias current (I_b)	300 μA					
Laser power (P_i)	100 μW			10 μW		
Laser wall-plug efficiency‡ (ξ)	25 %			25 %		
π phase shift voltage (V_π)	4 mV			4 mV		
Datarate (R)	1 GS/s			25 GS/s		
Injection power (P_{inj})	1 μW			1 μW		
Power EO modulation§ (P_m)	3.7 nW					
Power for optical energy¶ (E_{opt})	400 fJ/symbol	j=9×9	2.5 fJ	1.6 fJ/symbol	j=32×32	0.8 aJ
Nonlinear activation‖ ($E_{\rm NL}$)	(400 fJ/symbol)	(j=9×9)	(2.5 fJ)	1.6 fJ/symbol	j=32×32	(0.8 aJ)
ADC** ($E_{\rm ADC}$)	∼ 1 pJ/use [5]	i=28×28	0.5 fJ	∼ 1 pJ/use	i=10^6	0.5 aJ
Integrator ($E_{\rm INT}$)	∼ 1 fJ/use [6]	i=28×28	1 aJ	∼ 1 fJ/use	i=10^6	0.5 zJ
Energy of TIA** ($E_{\rm TIA}$)	∼ 1 pJ/use [7]	i=28×28	0.5 fJ	∼ 1 pJ/use	i=10^6	0.5 aJ
DAC†† ($E_{\rm DAC}$)	∼ 0.5 pJ/use [8]	j=9×9	3 fJ	∼ 1.6 aJ/use	j=32×32	0.8 zJ
Memory access‡‡ ($E_{\rm MEM}$)	∼ 100 fJ/access	j=9×9	0.6 fJ	∼ 100 fJ/access	j=32×32	50 aJ
Total energy (E_{tot})			∼7 fJ			∼50 aJ

*Fanout: j=9×9 and j = 32×32 are spatial fanout with phase mask. Data encoding to the laser field consumes energy for memory access and digital-to-analog converter (DAC). This energy cost is split to j times when the beam is fanned out for parallel operations. Similarly, the analog-to-digital converter (ADC), trans-impedance amplifier (TIA) and integrator are triggered once after integrating i time steps, so their operation rate is a factor of i slower than the clock rate, and their energy consumption per use is divided by the $2{\cdot}i$ operations. i=28×28 corresponds to time integration over an image of 28×28 pixels.

†OP: operation in multiply and accumulate (MAC). Every symbol in data encoding, and every use of ADC, TIA, integrator, DAC, and memory access computes 1 MAC=2 operation (OP).

‡The wall plug efficiency is $\xi = P_i/P_b = 25$ %. The electrical power for laser generation is determined by the product of the bias voltage and the current flowing through the VCSEL P_b=V_b×I_b=390 μW.

§The power used for electro-optic (EO) data modulation is P_m=$V_\pi^2/R_{\rm VCSEL}$=3.7 nW with the VCSEL resistance $R_{\rm VCSEL}$=V_b/I_b=4.3 kΩ. The corresponding energy per operation is E_m=P_m/R=3.7 aJ/sample with R =1 GS/s.

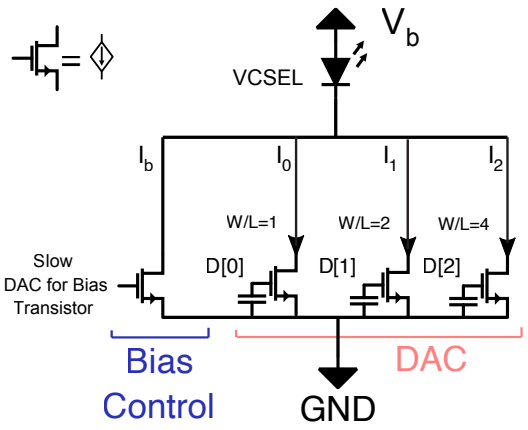
¶This includes the electrical power of laser generation, injection locking and data modulation, $E_{opt} = (P_b + P_{\rm INJ}/\xi + P_m)/(2R)$. Here the leader laser for injection locking is a VCSEL with the same wall-plug efficiency ξ. The power for injection locking and data modulation is negligible compared to that for laser generation. With future improvement with R=25 GS/s, i=10^6 and j=32×32, the laser may consume 40 μW electrical power to emit 10 μW optical power, which fans out to 10 nW (j=32×32) per homodyne detector. This power is sufficient for an SNR=100 in the VCSEL-ONN, the optical power consumption reaches ϵ_p=0.8 aJ/OP.

‖The homodyne nonlinearity consumes the same power as the optical energy.

** Standard CMOS ADC and TIA operates at 1 pJ/bit, corresponding to 0.5 pJ/OP. Our system performs signal amplification and ADC after integrating over i time steps. So the ADC and TIA energy costs are reduced by a factor of i.

†† Conventional CMOS digital-to-analog converters (DAC) operates at 0.5 pJ/use [8], corresponding to 0.25 pJ/OP. Each DAC generates a neuron activation that is fanned out to 9×9 copies, leading to 3 fJ/OP after fanout. This posts a limitation to the existing system. However, these CMOS DACs are designed to output at voltage >1 V [8], which is practically not required in our system because our VCSELs operate with V_π=4 mV. As E_{DAC} scales quadratically with operation voltage [9], it can be significantly reduced at low voltage operation. However, modern DACs are not designed to operate with such a low supply voltage. We propose a current-drive DAC circuit as Fig. S6, where the DAC supplies the power required for EO modulation and consumes only $E_{\rm DAC} = CV^2/2$=1.6 aJ/use, where C=200 fF/bit is the capacity for 1-mm wire, and V=0.4 mV is the V_π. This DAC uses a larger bias-control transistor to bias the VCSEL into lasing. Smaller transistors, sized such that the current they pull is staggered in powers of 2, convert a digital signal to a current that has the same effect as a 4 mV peak-to-peak drive voltage.

‡‡Another energy cost is for delivering the signal from memory to the DAC. This charges the electronic wires, with $E_{\rm MEM} = CV_{\rm MEM}^2/2$=100 fJ/access, where C=200 fF/bit is the capacity for 1-mm wire, and $V_{\rm MEM}$=1 V is the voltage for switching gates of transistors in electronic memory.



Supplementary Figure 6: Proposed optoelectronic design. Proposed CMOS circuit for driving VCSELs at ultralow V_π. The circuit consists of a bias controller for laser generation and a bias transistor for wavelength tuning in injection locking. The VCSEL is forward biased with a voltage of V_b and the current I_b flows to a large transistor to tune the VCSEL wavelength for injection locking. As the current of the transistor $I_{transistor} \propto W_T/L_T$, where W_T is the width and L_T is the length, the size of the DAC transistors is designed to be hundreds of times smaller than that of the bias transistor. The small DAC transistors pull currents from I_b, producing small-signal modulation for low V_π modulations.

IV. COMPARISON OF STATE-OF-THE-ART COMPUTING HARDWARE

.

A. Neural network models

Table III: Model size and accuracy of integrated-optics-related ONN-demonstrations

Ref.	dataset	Input data size	Model size	parameters in optical system	Experimental accuracy
Ashtiani 2022 [10]	4-letter EMNIST	3×4 pixels	3×4→3→2	40	89.9%
Xu 2021 [11]	MNIST	30×30 pixels	5×5×3 (conv layer) 72×10 (fully connected layer)	800	88%
Feldmann 2021 [12]	MNIST	28x28 pixels	4×2×2 (conv. layer)	16	95.3%
Shen 2017 [13]	Vowel	1×4	4×4	16	76.7%
This work	MNIST	28x28 pixels	28x28→100→10→10	79,410	93%

B. Energy consumption and footprint

The performance of digital computers and ONN systems is based on the survey of Ref. [14] and the analysis in Ref. [10], respectively.

Table IV: Performance of state-of-the-art computing hardware.

Hardware		compute density (TeraOP/(mm^2·s))	Energy efficiency (TeraOP/J)	Comments
Digital computer	Google TPU [15]	0.28	0.4	
	NVIDIA A100 [14, 16]	0.35	0.72	
	GraphCore IPU2 [14, 17]	0.17	1.0	
Optical Neural network	Photonic tensor core [18]	1.2	0.4	optical energy[d]
	Photonic deep neural network [10]	3.5	2.9[a]	optical energy
	Photonic deep neural network [10]	0.03[b]	0.07[c]	full-system energy[e]
	diffractive neural network [19]	not specified	0.7[f]	optical energy
	This work (now)	6	140	full-system energy
	This work (now)	25	400	optical energy
	This work (future improvement)	2000	20000	full-system energy

a. The energy efficiency is quoted 346 fJ/OP, which corresponds to 2.9 TeraOP/J.

b. The throughput in Ref. [10] is 0.27 TeraOP/s with a chip area of 9 mm^2, corresponding to a throughput density of 0.03 TeraOP/(mm^2·s).

c. The end-to-end energy efficiency in Ref. [10] is 14 pJ/OP, which is inverted to 0.07 TeraOP/J.

d. Optical energy consumption includes the power used for laser generation and driving EO modulators for data encoding.

e. Full system performance includes optical energy consumption, nonlinear activation, data readout, signal amplification, ADC, DAC, and memory access.

f. The energy efficiency quoted in [19] might be counted from 2-bit input precision (DMDs) and ineffective multiplications with the diffraction matrix (few trainable parameters)

Comparison between VCSEL-ONN and diffractive deep neural networks (D^{2}NNs) We compare the performance of our VCSEL-ONN to diffractive deep neural networks (D^{2}NNs). Similar to D^{2}NNs, our system based on

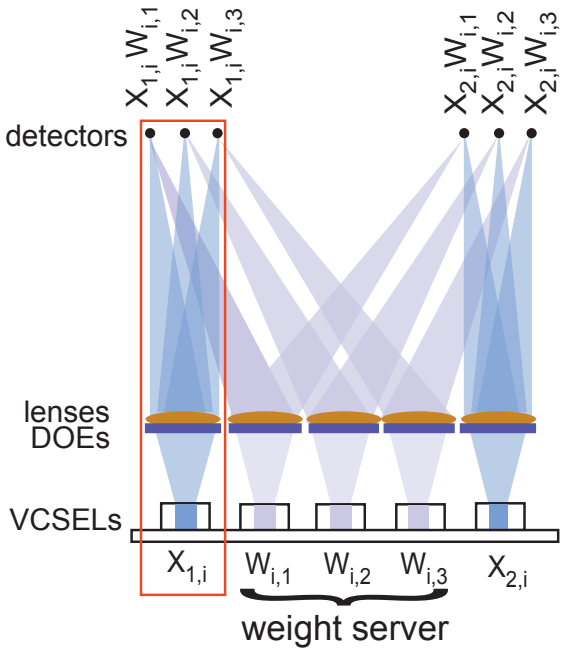
homodyne detection potentially allows complex number operations. Different from D^2NNs, our system operates matrix-vector multiplication with O(N) devices, which a diffractive neural network requires O(N2) devices: D layers, each layer with N pixels. The optical energy efficiency of 1.4 pJ/OP was quoted in a reconfigurable D^2NN [19], which is 500 times lower than our VCSEL-ONN. This is mainly due to the low frame of the digital mirror device and the spatial light modulators and the large pixel size. Moreover, existing D^2NN use detector nonlinearity, which requires high optical power to operate; a fast, low-power, scalable nonlinearity in D^2NNs is yet to developed.

Regardless of ONN architecture, compared to the digital mirror devices (DMD) and spatial light modulators (SLM) used in D^2NNs [19], the data rate of a commercial VCSEL with chip area ($<$0.01 mm^2) operating at 25 GS/s with 8 bits precision is the same as $>$100 DMDs in total, each operating with 4 million pixels (Mpx) at 10 kilo-frames per second (kfps) with 10 kfps$\times$4 Mpx/2^8=156 MS/s. The data rate of a single VCSEL is more than 60 SLMs in total, each operating at 200 fps with 2 Mpx, 200$\times$2 Mpx=400 MS/s. Each SLM pixel has 8-bit precision.

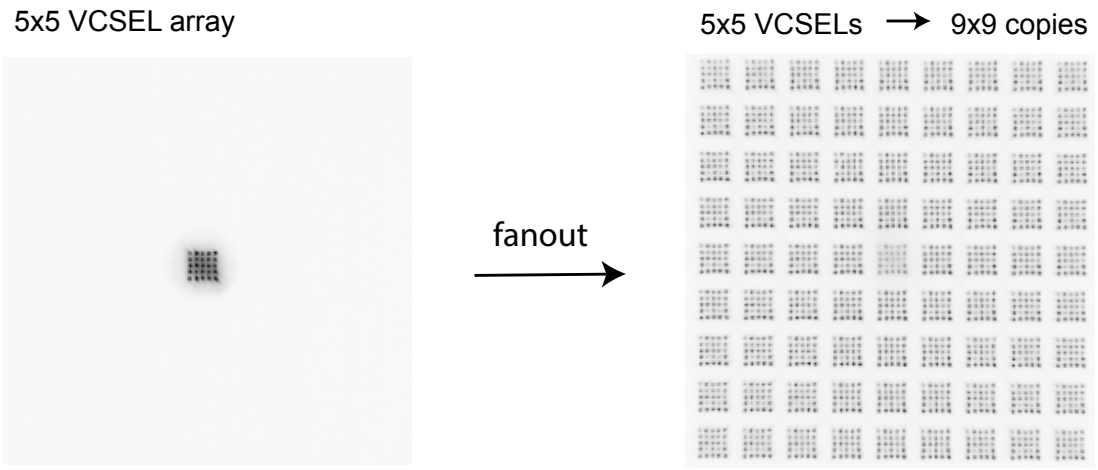
V. BROADCASTING THE WEIGHT SERVER

In the proposed scheme, the weight beams that encode the weight matrix [W_0, W_1, ..., W_j] can be fan-out to k copies, enabling to multiply with a batch of k input vectors simultaneously, enabling matrix-matrix multiplication $Y_{(k\times j)} = X_{(k\times i)}W_{(i\times j)}$. Fig. S7 shows a simple schematic for computing with k=2 input vectors and j=3 weight vectors, whereas i encoded in time steps is not shown here. The high parallelism allows operating matrix-matrix multiplication with $k + j$ VCSELs.

Shown in Figure S8 is the experimental result of the 5×5 VCSEL array (left) being fanned out onto a camera. The VCSELs are fanned out with a phase mask (MS-225-970-Y-A, Holoor.co) that splits the beam to 9x9 copies (right).



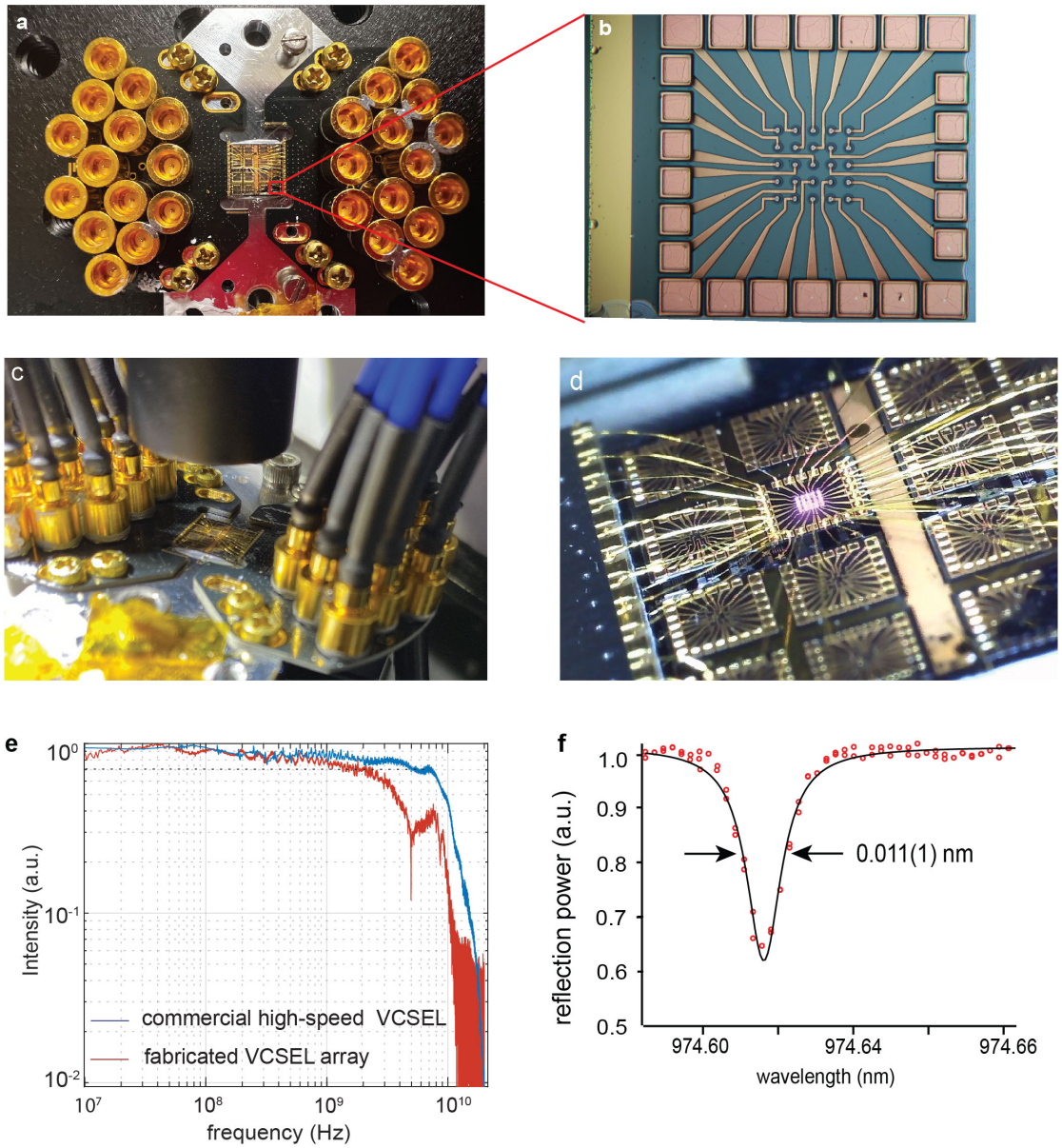
Supplementary Figure 7: Scheme of broadcasting weights for matrix-matrix multiplications. (a) The weight matrix $W_{i,j}$, generated with j weighting laser transmitters and i time steps in the weight server, is fanned out to multiply with k input VCSELs in parallel. Noting that the VCSEL array and the fanout are 2 dimensional and only the 1D plane is plotted. k=2, j=3 is shown in the example.



Supplementary Figure 8: Experimental fanout. The 5×5 VCSELs in an array is fanned out to k=9×9 copies, potentially enabling to process 81 input vectors simultaneously.

VI. VCSEL ARRAYS

One of our fabricated samples consisting of 16 arrays of 5×5 VCSELs (with 400 VCSELs in total) is shown in Fig. S9. We wire bond the arrays to a high-speed printed circuit board (PCB), which connects each individual VCSEL to an electronic driver using a bias tee. At the DC port of each bias tee, a single piece of AA battery (1.5 V) supplies all the VCSELs with a variable voltage divider on each channel. That allows individual voltage biasing to finely tune the wavelength of the VCSELs for injection locking. The AC port of the biased tee is connected to a high-speed arbitrary waveform generator (AWG) for data encoding. The output voltage of the AWGs is attenuated to 4 mV, matching to the V_π of the VCSELs.



Supplementary Figure 9: VCSEL samples. **a.** wire-bonded VCSEL arrays. The sample consists of 16 arrays, each with 5×5 VCSELs. The VCSELs are individually wired to an external co-axis cable using a PCB. **d.** High-resolution image with 5×5 VCSELs. **c.** The sample in the setup. The beams are collected with an achromatic lens. **d.** 5×5 wire-bonded VCSELs emits simultaneously. **e.** the 3 dB bandwidth of the fabricated VCSEL is 2 GHz. The bandwidth of a commercial VCSEL is measured 7 GHz. **f.** the linewidth of the on-chip VCSEL resonance is 0.011(1) nm, corresponding to a cavity Q factor of 10^5.

The bandwidth of the VCSELs is measured with a vector network analyzer. The 3 dB bandwidth of our fabricated VCSEL is 2 GHz. We scanned the cavity resonance with an external tunable laser and measured the reflection spectrum. The Lorentzian fit to the VCSEL resonance reveals a full-width-at-half-maximum of $\delta\nu$=0.011(1) nm, corresponding to 3 GHz. The Q-factor of the VCSEL cavity is Q=$\nu/\delta\nu$=10^5. The measured VCSEL bandwidth is 67 % of the photon-lifetime limit. The discrepancy might be due to the cavity dynamics modified by the laser emission (the linewidth scan is performed below lasing threshold, while the modulation bandwidth is measured above lasing threshold). To explore the potential of the VCSEL platform, we employ a commercial VCSEL [20] is about 7 GHz and it was used to showcase the data encoding at 25 GS/s.

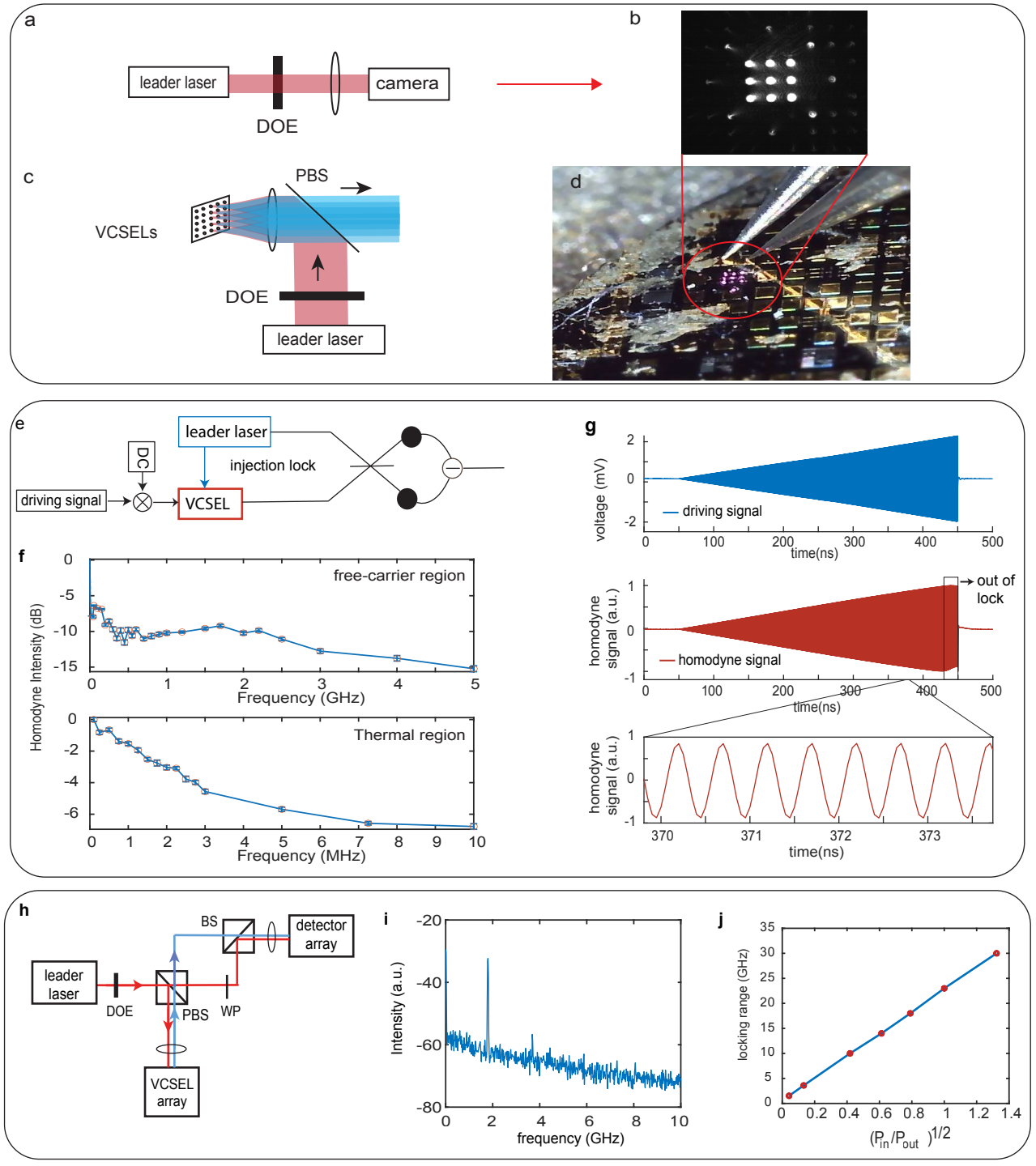
VII. INJECTION LOCKING

The setup of injection-locking is shown in Fig. S10. Here the sample is with arrays of 3×3 VCSELs. In **a**, the leader laser passes through a diffractive optical element (DOE, model MS-711-K-Y-X, HOLOOR.co) is fanned out to a 3×3 beam array with a separation angle of 0.26 degrees in the Fourier plane (Fig. S10b). The focus length of the lens is chosen 17.5 mm to provide a separation distance of 80 μm between the beam spots, matching the pitch of our VCSEL sample Fig. S10d.

We perform interferometric detection to characterize the injection locking of the VCSEL array with an experimental setup in Fig. S10. The leader laser passing through the DOE is split into two parts by the polarizing beam splitter. One part is used for injection locking (Fig. S10), and the other part passing is overlapped to the beams of the VCSELs using a beamsplitter. We couple the combined beams (one VCSEL and a copy of the leader laser) to a 4×4 fiber array of 160-mm pitch (twice as that of the VCSELs) with a focus lens of 35-mm focus length. The beatnote between the leader laser and a VCSEL is detected, as shown in Fig. 10b. We measured the injection locking range as a function of injection power in Fig. 10c by monitoring the beatnote with a high-speed photodetector (bandwidth 12 GHz) and a spectrim analyzer. The beatnote shifts to DC frequency when the VCSEL falls within the injection locking range. We tuned frequency of the leader laser while reading out the detuning using a wavemeter. The observed injection locking range is proportional to the square root of the input power.

We characterize the frequency response of the injection-clocked VCSEL with homodyne balanced detection (Fig. 10). The leader laser and the injection-locked VCSEL is overlapped with a beamsplitter and a balance detector records the homodyne interference. By applying a sine wave increasing frequencies (and constant amplitude), we observe an amplitude decay in the homodyne signal as a function of frequency (10b). The result reveals that the frequency response of our injection-lock VCSELs in the free-carrier region (f>10 MHz) is 10 times weaker than the thermal region (f<1 MHz). Similar frequency response has been observed in Ref. [21, 22].

The π phase shift voltage is measured by applying a sine wave with increasing voltage using the data encoding scheme of (Fig. S10a). As shown in Fig. S10c, the driving voltage (blue) increases from 0 to 4.2 mV (in x-axis from 50 ns to 450 ns). The homodyne signal increases linearly with the driving voltage (red) until the VCSEL falls out of injection lock at the peak-to-peak voltage of about 4 mV, which suggests a V_π=4 mV. The envelope of this homodyne signal is plotted as a function of driving voltages in the main text Fig. 2h.



Supplementary Figure 10: Injection locking. **a-d**. Example of injection locking over a VCSEL array. DOE: diffractive optical element. PBS: polarizing beam splitter. **e-g** Injection locking range. **e.** experimental setup for characterization. DOE: diffractive optical element. PBS: polarising beam splitter. WP: half waveplate. BS: beam splitter. **f.** beatnote of the leader laser and a VCSEL at 1.9 GHz, out of injection lock range. **g** the injection lock range is proportional to the square root of the input laser power. The VCSEL output power is fixed 100 μW. The error bar indicates the maximum and the minimum of each data value from 10 consecutive measurements. The data points are presented as mean values of the 10 measurements. **h-j** Homodyne detection of injection locking bandwidth and V_π. **h.** Setup of homodyne detection. **i.** frequency response of injection-locked VCSEL. **j.** experimental characterization of π phase shift voltage at 2 GHz.

VIII. DATA MODULATION AND DEMODULATION

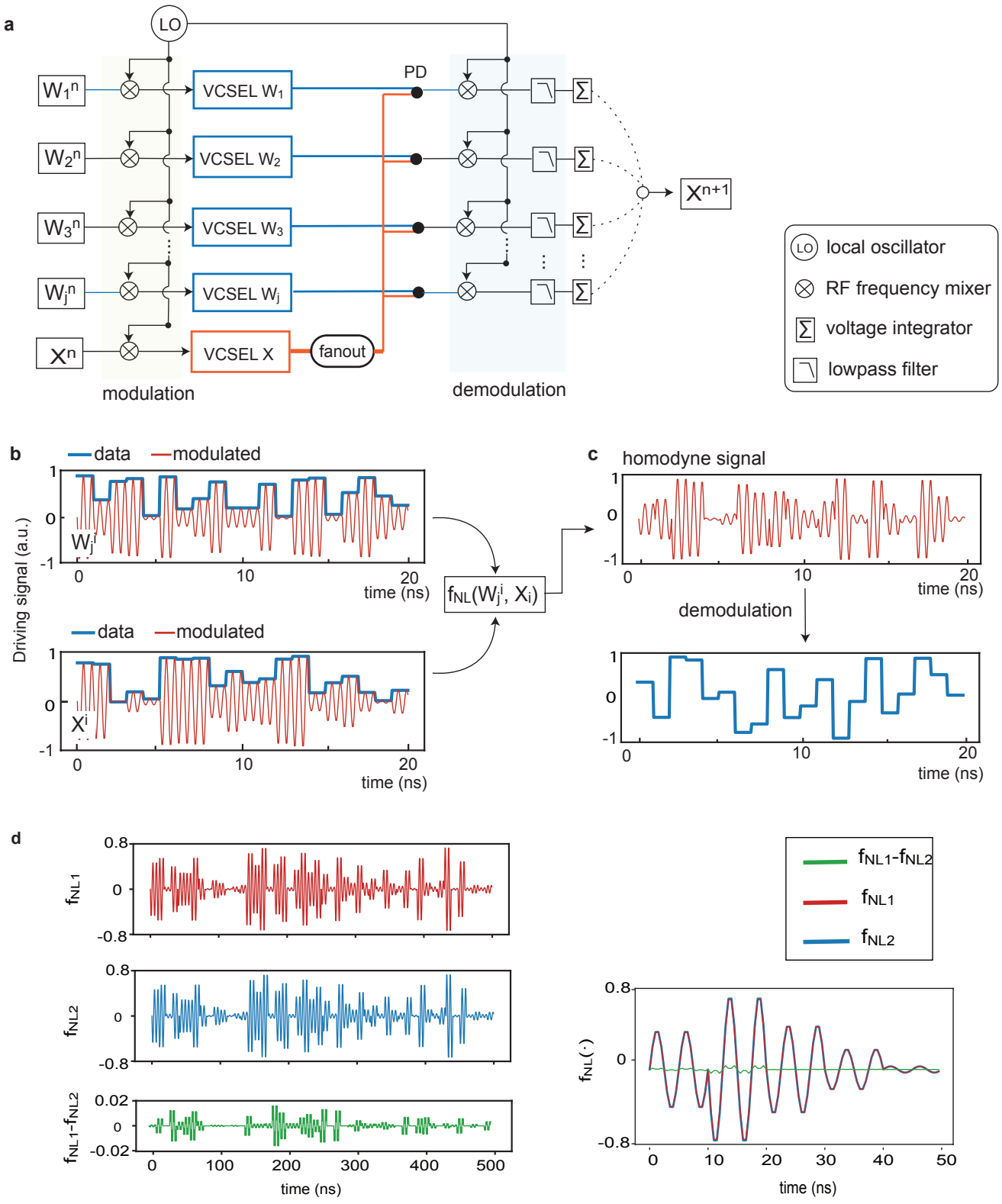
The thermal response on our injection-locked VCSELs leads to slow phase drifts in the homodyne signal at 1 GS/s speed modulation. To decouple the thermal effect, we encode the data (at 1 GS/s) with a fast local oscillator at $\omega_{\mathrm{LO}} = 2\pi{\cdot}2$ GHz using a frequency mixer, as shown in Fig. S11. The modulated driving signal (Fig. 11b) averages at 0, which is good for thermal balance in every time step. The generated homodyne signal is represented as

$$f_{NL1}(W_{ij}, X_i) = W_{ij}\sin(\omega_{LO})\sqrt{1 - X_i^2\sin^2(\omega_{LO})} - X_i\sin(\omega_{LO})\sqrt{1 - W_{ij}^2\sin^2(\omega_{LO})} \tag{14}$$

To confirm that the signal can be demodulated correctly. We compare the simulation result of Eq. 14 to that of f_{NL2} in Fig. S11

$$f_{NL2}(W_{ij}, X_i) = \left[W_{ij}\sqrt{1 - X_i^2} - X_i\sqrt{1 - W_{ij}^2)}\right]\sin(\omega_{LO}), \tag{15}$$

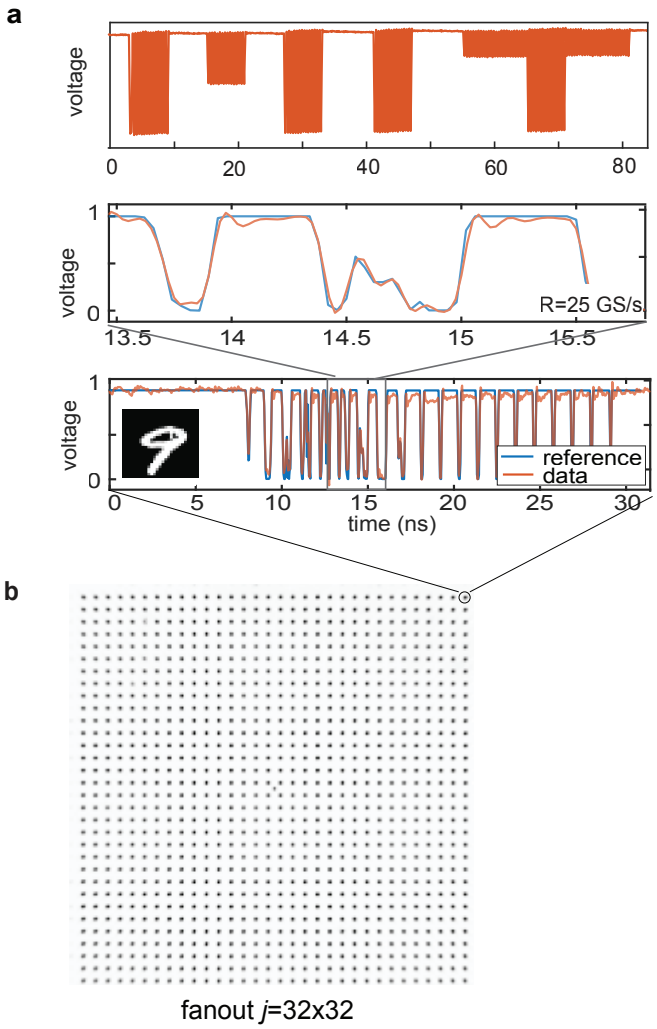
where the sine wave can be demodulated to produce an accurate homodyne product.



Supplementary Figure 11: Data modulation and demodulation. (a) The data modulation scheme in the experimental setup. A local oscillator is used for data modulation and demodulation. LO: local oscillator. PD: photo-detector. **b.** Modulated data. the input and weight data are encoded by mixing with a local oscillator at the frequency of two times the data rate. **c.** The generated homodyne signal is demodulated with the local oscillator to retrieve the compute result. **d** Verification of compute fidelity using data modulation scheme. The red curve is computed with f_{NL1} and the blue is from f_{NL2}. The $f_{NL1} - f_{NL2}$ residual is 0.7 % over 10,000 data points, which is negligible.

IX. LARGE FANOUT FACTOR AND HIGH-SPEED MODULATIONS

A commercial VCSEL (VI System V25-1550) with a nominated 3dB bandwidth of >13 GHz is used to benchmark the high speed data modulation at 25 GS/s (Fig. S12a). The VCSEL beam is fanned out to $j = 32 \times 32$ copies, potentially enabling computing at rates up to 50 TetaOP/s.



Supplementary Figure 12: A VCSEL with data encoding at 25 GS/s is spatially fanned out to j=32×32 copies.

[1] N. R. Newbury, I. Coddington, and W. Swann, Opt. Express **18**, 7929 (2010).

[2] T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon, Nature Communications **13**, 1 (2022).

[3] A. Sludds, S. Bandyopadhyay, Z. Chen, Z. Zhong, J. Cochrane, L. Bernstein, D. Bunandar, P. B. Dixon, S. A. Hamilton, M. Streshinsky, A. Novack, T. Baehr-Jones, M. Hochberg, M. Ghobadi, R. Hamerly, and D. Englund, Science **378**, 270 (2022), https://www.science.org/doi/pdf/10.1126/science.abq8271.

[4] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, The Journal of Machine Learning Research **18**, 6869 (2017).

[5] B. M. Pietro Caragiulo, Clayton Daigle, Dac performance survey 1996-2020 (2022).

[6] E. Yang and T. Lehmann, ISCAS 2019 (2019).

[7] E. Yang and T. Lehmann, in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)* (2019) pp. 1–5.

[8] O. Wada, T. T. Ta, S. Tanifuji, S. Kameda, N. Suematsu, T. Takagi, and K. Tsubouchi, in *2012 Asia Pacific Microwave Conference Proceedings* (2012) pp. 1118–1120.

[9] O. Morales Chacón, J. J. Wikner, C. Svensson, L. Siek, and A. Alvandpour, Analog Integrated Circuits and Signal Processing **111**, 339 (2022).

[10] F. Ashtiani, A. J. Geers, and F. Aflatouni, Nature **606**, 501 (2022).

[11] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss, Nature **589**, 44 (2021).

[12] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, Nature **589**, 52 (2021).

[13] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, *et al.*, Nature Photonics **11**, 441 (2017).

[14] M. Khairy, Tpu vs gpu vs cerebras vs graphcore: A fair comparison between ml hardware (accessed on June 13 2022).

[15] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, *et al.*, in *Proceedings of the 44th annual international symposium on computer architecture* (2017) pp. 1–12.

[16] NVIDIA, The universal system for ai infrastructure (accessed on June 13 2022).

[17] GraphCore, Introducing 2nd generation ipu systems for ai at scale (accessed on June 13 2022).

[18] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, Nature **569**, 208 (2019).

[19] T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, and Q. Dai, Nature Photonics **15**, 367 (2021).

[20] VI Systems GmbH, Up to 25 gbit/s vcsel single mode fiber-coupled module (1550 nm) (accessed on June 13 2022).

[21] S. Bhooplapur, N. Hoghooghi, and P. J. Delfyett, Opt. Lett. **36**, 1887 (2011).

[22] S. Kobayashi and T. Kimura, IEEE Journal of Quantum Electronics **18**, 1662 (1982).