nature sensors

Supplementary information

https://doi.org/10.1038/s44460-026-00123-2

Fully in-memory photonic computing

In the format provided by the authors and unedited

Supplementary information for **Fully in-memory photonic computing**

Supplementary Note-1. Theory of fully in-memory neuron 2
Supplementary Note-2. FLARE memory neuron model 3
Supplementary Note-3. FLARE chip packaging 5
Supplementary Note-4: Relation of FLARE to existing technologies. 6
Supplementary Note-5: Benchmarking the performances of FLARE against that of electronic CNNs. 9
Supplementary Note-6: System performance of fully in-memory computing 9
Supplementary Table 1. Performance benchmarking. 15
Supplementary Table 2. Performance Summary. 16

Supplementary Note-1. Theory of fully in-memory neuron

1.1 Cascadable neuron.

The dynamics of the photonic and electronic cavities are described by

$$\frac{d}{dt}o = (j\omega_o - \gamma_o)o + \sqrt{2}s_o \tag{S1}$$

$$\frac{d}{dt}e = (j\omega_e - \gamma_e)e + \sqrt{2r_e}s_e. \tag{S2}$$

In the main text, the coupling coefficient r's are omitted for clarity without loss of generality. By introducing a coupled cavity, the photonic and electronic fields become interconnected through their respective loss terms:

$$\gamma_o = \alpha_{eo}\, e \text{ and } \gamma_e = \alpha_{oe}\, o. \tag{S3}$$

To realize a cascadable neuron, we consider the k-th layer o_k, with signal propagation proceeding as $o_k \to e \to o_{k+1}$. The first conversion $o_k \to e$ is mediated via optoelectronic coupling, where $\gamma_e = \alpha_{oe}\, o_k$. Substituting γ_e into the Eq. S2 yields $\frac{d}{dt}e = (j\omega_e - \alpha_{oe}\, o_k)e + \sqrt{2r_e}s_e$. Given the inherent parameters ω_e, and r_e, the electronic field e is parameterized jointly by the photonic input o_k and the electronic injection s_e, leading to the representation of $e(s_e, o_k)$. The $(k+1)$-th photonic layer then evolves according to $\frac{d}{dt}o_{k+1} = (j\omega_o - \alpha_{eo}\, e(s_e, o_k))o_{k+1} + \sqrt{2r_o}s_o$. This formulation establishes a direct, fully photonic inter-layer connection, eliminating the need for intermediate digital retrieval and thereby enabling cascadability across layers.

1.2 Long-short term memory.

We now analyze the solution $e(s_e, o_k)$ under simultaneous electronic injection s_e and photonic input o_k, described by $\frac{d}{dt}e = (j\omega_e - \alpha_{oe}\, o_k)e + \sqrt{2r_e}s_e$. Including the intrinsic RC decay of the electronic circuit with time constant τ_e, this becomes

$$\frac{d}{dt}e = \left(j\omega_e - \alpha_{oe}\, o_k - \frac{1}{\tau_e}\right)e + \sqrt{2r_e}s_e \tag{S4}$$

When $\alpha_{oe}\, o_k \ll \frac{1}{\tau_e}$, the intrinsic electronic decay dominates, yielding a memory time $\tau_e = RC$. Conversely, when $\alpha_{oe}\, o_k \gg \frac{1}{\tau_e}$, the memory is dominated by photonic coupling with a characteristic timescale $\tau_o = 1/\alpha_{oe}\, o_k$. Experimentally, long- and short-term dynamics are separated such that $\gamma_o = \alpha_{eo-L}\, e_L + \alpha_{eo-S}\, e_S$, where L and S stand for the long- and short-term memory, respectively. The corresponding electronic dynamics are given by,

$$\frac{d}{dt}e_L = \left(j\omega_e - \frac{1}{\tau_e}\right)e_L + \sqrt{2r_e}s_{e-L} \tag{S5}$$

$$\frac{d}{dt}e_S = (j\omega_e - \alpha_{oe}\, o_k)e_S + \sqrt{2r_e}s_{e-S} \tag{S6}$$

Separation of the long- and short- term memory is achieved by capacitor design. The distinct dynamics between the long-and short-term dynamics are demonstrated in Figs. 3 and 4.

More specifically, the LT-Mem dynamics e_L is governed by the resistor-capacitor of the reverse-biased junction and the parallel capacitor. Fig. S3a shows the I(current)-V(voltage) characteristics of the junction by sweeping the bias voltage from -6 V to 1 V. The results show that when reverse-biased from -6 V to -0.5V, the measured current ranges from 30 fA to 75 fA, corresponding to an equivalent resistance of approximately 16.7 TΩ to 75.9 TΩ. Given a junction area of 0.22 μm × 40 μm = 8.8 μm², these measurements indicate a current density in the range of 3.4–8.5 fA/μm². Fig. S3c shows the capacitance of the FLARE long-term memory is approximately 380fF. The corresponding time constants calculated based on the first principles t = RC are visualized in the Fig. S3d and span a decay time from 6.39s to 28s, which aligns well with the 7.45-s time constant measured in the experiments.

The ST-Mem dynamic e_S is governed by the detector junction capacitance C_{ST}, the input capacitance C_{iss}, and the open-circuit resistance R_{ds} of the MOSFET, as illustrated in Fig. S4, where the equivalent circuit diagram of the short-term memory cell is visualized. Parasitic capacitances associated with the switching transistor, including C_{ds}, C_{gs}, and C_{gd}, are also considered. The ST-Mem operation consists of four stages: switch on, capacitor charging, discharging, and switch off. The three processes as illustrated in Fig. S4b, including switch on, capacitor charging, and switch off collectively determine the operation frequency, while the discharge speed governed by the incident optical power is sufficiently fast for bandwidth measurements. Accordingly, the total frequency is given by $f_{st} = 1/(2\pi R_g C_{iss} + 2\pi R_{ds} C_{ST}/2)$. Considering that the measured C_{iss} is on the order of 200 fF (Fig. S4c), R_g of 50 Ω, R_{ds} of 1.25 kΩ, and C_{ST} of 35 fF, the theoretical frequency averages to 5.04 GHz (Fig. S4d), which agrees with the experimental measurements.

Supplementary Note-2. FLARE memory neuron model

Following the design illustrated in Eq. S5 and Eq. S6, we developed the differentiable network model for the FLARE neuron, enabling its optimization and training. The modeling process of the FLARE neuron is divided into two parts: the circuit part and the resonator part.

In the first part, the network models the voltage variation across the FLARE neuron when activated by input light. Once the FLARE neuron parameters are set and the transistor switch is turned off, the circuit can be abstracted as a parallel combination of three components: a capacitor, a resonator, and a photodiode, as shown in Extended Data Fig. 8. The resonator is modeled as an equivalent resistor, while the photodiode is represented by a parallel combination of a reverse-biased diode and a constant current source. This current source generates a photocurrent I_0, which is proportional to the input light intensity. Extended Data Fig. 8 shows the layout of FLARE neuron and the abstract schematic of the circuit when the switch is off.

When the switch opens, the voltage across the capacitor initially remains unchanged and subsequently discharges through the resonator and the photodiode until it is clipped at zero by the diode. This circuit behaves as a first-order dynamic circuit

system, and the voltage $u(t)$ across the capacitor can be described as:

$$u(t) = \begin{cases} -I_0R + (V_0 + I_0R)e^{-\frac{t}{\tau}}, & t \le \tau \ln\left(\frac{V_0}{I_0R} + 1\right) \\ 0, & t > \tau \ln\left(\frac{V_0}{I_0R} + 1\right) \end{cases} \quad \text{(S7)}$$

where V_0 is the steady-state voltage across the capacitor when the switch is on, I_0 is the photocurrent generated by the input light, $\tau = RC$ is the time constant of the circuit, R is the equivalent resistance of the resonator, and C is the capacitance of the unit.

In the FLARE neuron, time constants of the long- and short-term memory τ are designed on the order of second and nanosecond, respectively while the integration time t is set in the regime $t \ll \tau$. This allows the exponential term to be approximated using its first-order Taylor expansion, $e^x \approx x + 1$ for $|x| \ll 1$, giving

$$u(t) = \begin{cases} V_0 - (V_0 + I_0R)\frac{t}{\tau}, & t \ll \tau \text{ and } t \le \tau \ln\left(\frac{V_0}{I_0R} + 1\right) \\ 0, & t \ll \tau \text{ and } t > \tau \ln\left(\frac{V_0}{I_0R} + 1\right) \end{cases} \quad \text{(S8)}$$

Applying an additional first-order approximation for the logarithm $\ln(1 + x) \approx x$ for $|x| \ll 1$, further simplifies the expression to

$$u(t) = \max\left(0, V_0 - (V_0 + I_0R)\frac{t}{\tau}\right), \ t \ll \tau \quad \text{(S9)}$$

Thus, for a given integration time t, the output voltage $u(t)$ depends only on the initial bias voltage V_0 and the photocurrent I_0, which is directly proportional to the input optical intensity I.

In the experiment, we calibrated the output optical intensity $I(t)$ of the unit under various initial voltages V_0 and normalized input light intensities I. By further mapping the measured capacitor voltage to the corresponding output optical intensity, we constructed a calibration matrix of the outputs as a function of input optical intensity and initial bias voltage V_0, which is expressed as $u(t) = f(I, V_0)$. The network employs this matrix to predict the final capacitor voltage via two-dimensional linear interpolation. An example calibration result is shown in Fig. S1a.

The second part of the neuron models the influence of voltage variation on the optical transmission of the resonator. Fig. S1b depicts the relationship between the transmission of a FLARE neuron and the normalized bias voltage. As the voltage across the resonator increases, the resonance peak undergoes a redshift (shift to longer wavelengths) accompanied by an increase in resonance depth.

During calibration, the transmission spectrum of the resonators under different voltages in a steady-state condition were characterized. This relationship can be expressed as:

$$T = T(\lambda, u')$$

where T represent the transmission, λ is the wavelength of the probe light, and u' is the normalized bias voltage of the unit. Due to limited sampling points, the following procedure were used to fit the transmission T as a function of λ and u':

1. Measure the transmission at probe wavelengths $\lambda_1, \lambda_2, \dots, \lambda_n$ under normalized bias voltages $u'_1, u'_2, \dots, u'_m$, constructing a transmission matrix $T_{m\times n}$.
2. Align each row i of matrix by shifting its elements so that the resonance wavelength λ_{peak_i} is centered. The relationship between the resulting aligned matrix $T'_{m\times n}$ and the original matrix $T_{m\times n}$ can be described as:

$$t'_{i,j} = t_{i,j-peak_i+\lfloor n/2 \rfloor}, \qquad i = 1,2,\dots,m; j = 1,2,\dots,n$$

where $t_{i,j}$ and $t'_{i,j}$ represent the elements in row i and column j in the original and aligned matrices, respectively. Here, $peak_i$ denotes the index of the maximum value in row i of the original matrix $T_{m\times n}$. Any gaps created during the shifting process are filled using boundary values.

3. Perform two-dimensional linear interpolation on the aligned transmission matrix T' to obtain a multivariate transmission function of wavelength offset and normalized voltage $T' = f(\lambda - \lambda_{peak}, u')$.
4. Perform one-dimensional linear interpolation to map normalized bias voltage u' to the resonance peak wavelength $\lambda_{peak} = g(u')$.

The final interpolated function of transmission can be expressed as

$$T(\lambda, u') = f(\lambda - g(u'), u')$$

where $g(u')$ describes the linear interpolation between normalized bias voltage u' and the resonance peak wavelength.

Compared to direct interpolation of the original transmission matrix T, the calibration method described above provides a more accurate approach by accounting for the actual pattern of the single-peak transmission curve. The schematic diagram of the above method is shown in Fig. S1c.

Together, these two parts form a complete model of the FLARE neuron. The first part determines the capacitor voltage as a function of input light intensity and time, while the second part translates this voltage into the optical transmission spectrum. The resulting output serves as the input to the subsequent layers.

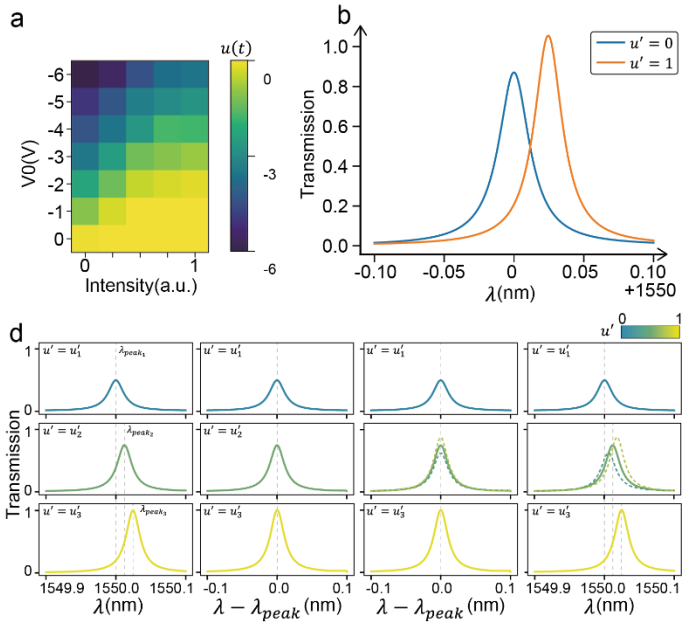

Fig. S1. Unit model of a FLARE neuron. a, Circuit calibration. Two-dimensional calibration matrix of output voltage $u(t)$ as a function of input optical intensity I and initial bias voltage V_0. **b,** Transmission response. Increasing voltage induces a redshift and deeper resonance. **c,** Resonance modeling. Procedure includes resonance peak alignment, two-dimensional interpolation, and offset restoration. Scale bar: 1 mm.

Supplementary Note-3. FLARE chip packaging

Electrical Packaging. Printed circuit boards (PCBs) were designed and manufactured to serve as the interface for the chip. Electrical connections between the chip bond pads and PCB pins were established using gold wire bonding, enabling parallel addressing

of the long-term memory cells on each chip. Several multichannel connectors (Samtec ERF8-040-05.0) were mounted on the bottom layer of the PCB to enhance electrical connectivity. A high-speed global control interface was designed to address the short-term memory units, which minimized gold wire lengths and reduce parasitic inductance. It also utilized a grounded coplanar waveguide structure to maintain a transmission line impedance of 50 ohms. Additionally, a grounding structure and shielding vias were incorporated between channels to ensure electromagnetic shielding. The global control signal was subsequently interfaced through side-mounted high-speed connector (Gwave SMA-KHD7). The assembly of the PCB and metal substrate was secured using M2.5 screws. The NTC thermistor was then affixed to the metal substrate with silver epoxy, finalizing the die bonding process after curing the epoxy at 110 degrees Celsius.
Optical Packaging. For optical signal transmission of three-dimensional computing cores (core-1 and core-2), a single-mode-fiber was employed with mode profile matched to the edge waveguide coupler on chip. For two-dimensional computing cores (core-3), two 24-channel fiber arrays were employed for the on-chip light coupling and readout. The alignment was performed by maximizing the optical transmittance between the waveguide and on-chip grating coupler aids, ensuring precise coupling between the memory chip and the optical fiber through motorized 6-axis stages. The measured alignment loss between the chip and the fiber was 3.1 dB per facet, and the fiber was securely attached to the chip with UV curing. Packaging was conducted using a 1550-nm light source. Exemplary packaged cores are displayed in Extended Data Fig. 4a.

Supplementary Note-4: Relation of FLARE to existing technologies.

Active-matrix architecture. The active-matrix architecture, which has been widely adopted in the display industry, reduces control complexity from $O(N^2)$ to $O(N)$. In contrast, due to the lack of an active-matrix architecture, photonic computing chips have been constrained by $O(N^2)$ control complexity. From a technical perspective, the limitation of realizing active control matrix on photonic integrated circuits arises from three main factors: (1) the absence of memory-enabled neuron architectures capable of retaining driving states, (2) crosstalk introduced by monolithic optoelectronic integration that compromises independent neuron addressing, and (3) the lack of a complete device design and fabrication pipeline that supports the full fabrication flow required for large-scale integration. In response to these challenges, we introduce several specific innovations in FLARE.

1. Fully in-memory neuron. The cornerstone of active-matrix driving is the ability of each neuron to retain its driving state. In conventional electronic systems, this functionality is provided by local digital memory circuits. However, photonic integrated circuits are typically optimized for ultrafast dynamics, leading to rapid decay of driving states. While phase-change materials have been explored for in-memory photonic computing, their high driving power and limited switching speed are incompatible with standard active-matrix driving schemes. In this work, we develop a coupled-resonator FLARE neuron that can be driven by electronic signals while

retaining its state over extended durations. This capability establishes the fundamental requirement for active-matrix control in photonic integrated circuits.

2. Optoelectronic neuron isolation. Another critical challenge is crosstalk induced by monolithic optoelectronic integration, which prevents independent neuron operation. As illustrated in Fig. S5, when a long-term memory neuron is addressed through a signal line, the driving signal can leak into short-term memory cells and perturb the bias states of neighboring neurons through both electrical interconnects and optical waveguides. To address this issue, we introduce an isolation module based on interleaved PN doping. This design effectively suppresses electrical and optical crosstalk, enabling truly independent addressing and operation of each FLARE neuron.

3. Monolithic fabrication. To physically implement the proposed parallel driving architecture, we develop a holistic fabrication process on a standard silicon photonics platform that supports the monolithic integration of electronic and photonic components. The resulting component library includes electronic and optoelectronic building blocks that are absent from conventional silicon photonics platforms, thereby enabling the realization of active-matrix circuits. Importantly, this fabrication workflow supports both the coupled-resonator FLARE neuron architecture and the neuron isolation strategy introduced in this work, making large-scale in-memory photonic computing systems practically realizable.

In summary, we present an optoelectronic coupled-resonator-based photonic in-memory neuron capable of long-term state retention, we propose neuron isolation techniques that enable large-scale independent neuron control, and we develop a fully integrated fabrication procedure to implement the complete architecture. Building up on these innovations, our work demonstrates the photonic computing chip with large-scale active-matrix control.

Sample-and-hold control. A sample-and-hold approach was developed for the multiplexed driving of photonic integrated circuits[1], which interrogate the thermal modulators by cointegrating the photonic integrated circuits with the electronic controllers. However, the referenced work is distinct from the FLARE in the scope and capability, especially in computational architecture, energy efficiency, processing speed, and reconfigurability.

Most importantly, the referenced work presents photonic driving circuits intended for controlling photonic integrated components, rather than neurons for computation. As such, it lacks essential computational elements, including layered computational circuits that map inputs to outputs, large-scale parallel optical neurons, and intrinsic nonlinearities required to support deep neural network architectures.

Specifically, the modulators are driven by external power supply, leading to continuous power dissipation with mA driving currents. Also, the driving signals are supplied by constant sources, restricting the modulation states to binary levels. The referenced architecture relies on thermo-optic modulators, inherently limiting the response speed.

We further elaborate on how the fully in-memory computing architecture of FLARE

advances neuromorphic computing by pushing the boundaries of neuron scalability, cascadability, and perception capability.

Scalability of memory neuron

The integrated control schematics of FLARE enable the large-scale, parallel reconfiguration of photonic neuron arrays, while simultaneously retaining the advantages of high speed and energy efficiency. Existing photonic in-memory computing methods generally exploit two distinct approaches for configuring memory neurons. The first approach employs all-optical laser pulses to thermally induce phase changes in materials. This method requires repeatedly translating optical fibers and carefully coupling light into the memory units, which makes it challenging to reprogram neuron arrays rapidly and at scale. The second approach uses electronic heating, which offers a more compact and efficient means of programming neurons. However, in existing reconfigurable photonic methods, every memory unit arranged in the two-dimensional plane must be routed to the one-dimensional edge. This requirement implies that $O(N^2)$ wires must be packed within a spatial footprint of only $O(N)$, which quickly becomes infeasible as the number of neurons increases. Due to this wiring bottleneck, point-wise control has an upper scalability limit of approximately 8×8. In FLARE we propose an opto-electronic mechanism for photonic memory that avoids slow thermal heating. More importantly, FLARE overcomes the scale limit by adopting a parallel control architecture realized by monolithic fabrication, addressing an $O(N^2)$ array using only $O(N)$ electrical interconnects. By selectively activating or deactivating column switches, the system can precisely control the weight-loading configuration of each individual unit within the array.

Neuron cascadability and nonlinearity

The ability to directly cascade signals between neurons is essential for constructing deep neural networks. However, current optoelectronic and electronic neuromorphic computing systems primarily focus on optimizing intra-layer computation, while inter-neuron connections continue to rely heavily on external digital memory. In most existing methods, the analog output of one neuron is detected, converted into the digital domain, buffered in digital memory, and then used to drive the neurons in the following layer. This dependency on analog-to-digital conversions and intermediate memory access introduces overhead, creating bandwidth bottlenecks that degrade overall throughput and energy efficiency. Emerging approaches attempt to overcome this bottleneck by performing computation in the analog domain. One possible solution is to employ amplification circuits to connect neurons directly. Yet, this approach requires high-power, bandwidth-limited, and structurally complex electronic circuits, which undermine the inherent energy efficiency of optical computing to TOPS/W level (picojoule-level per operation). FLARE fundamentally resolves these limitations by seamlessly integrating a photonic cavity with an electronic cavity. Through strong optoelectronic coupling, incident light injected into a neuron directly modulates the output light field. This design circumvents the bandwidth limitations imposed by intermediate analog-to-digital conversions and memory buffering. As a result, FLARE supports direct multilayer interconnections with ultrafast computation, endowing the system with the scalability and efficiency required to address large and complex computational tasks.

Perception

Perception represents one of the essential ingredients for advancing neuromorphic systems toward human-like intelligence. However, integrating sensing and deep neural network computation in a manner analogous to the human brain remains a substantial

challenge. One direction of research is the development of smart sensors, which redistribute the photocurrent generated by sensor pixels to implement a shallow neural network at or near the sensor. Although this approach allows local preprocessing, it supports only a single-layer network, while the computationally demanding deep neural network layers must still be executed on conventional digital processors after signal detection. In photonic computing, another strategy involves sensing coherent laser illumination and performing computation on-chip after optical coupling. However, the requirement of a laser light excludes the applicability of the general naturalistic scene featuring incoherent and partial coherent sources. The architecture of FLARE overcomes these constraints by embedding the photosensitivity directly within each memory neuron cavity. This design allows neurons to sense general light intensity, irrespective of coherence. Furthermore, thanks to the photonic cascadability of the system, this sensing capability can be directly integrated with deep neural network computation within a unified architecture.

Supplementary Note-5: Benchmarking the performances of FLARE against that of electronic CNNs.

For a more direct comparison, we evaluate the energy consumption of FLARE and electronic CNNs to perform the same tasks. Specifically, we trained a series of CNNs with parameter counts ranging from 0 to 40 K on the same task. Each CNN consists of three convolutional layers with max-pooling and residual connections, followed by a linear readout layer, with ReLU activation functions applied between the convolution layers. By varying the channel number in the hidden convolution layers, a set of CNNs with different parameter was constructed for the drone navigation task. Fig. S6 presents the mean squared error and energy consumption achieved by CNNs with different channel numbers. The energy consumption is also evaluated on an Nvidia H100 GPU operating at peak utilization by using the largest possible test batch size. For comparison, the FLARE network used in the task shown in Fig. 5 contains 99.33 K parameters and achieves a mean squared loss of 0.0329. To reach the same level of accuracy, FLARE requires only 18.6 nJ energy during inference while the Nvidia takes 4.0 mJ, indicating 2.15×10^5 times gain of efficiency when undertaking the same task.

FLARE continues to demonstrate advantages in terms of speed and energy consumption, even when compared to electronic convolutional networks.

Supplementary Note-6: System performance of fully in-memory computing

FLARE alleviates the computing bottlenecks of signal conversion and power-exhausting interface by performing all computations fully within in the memory. This holistic design not only reduces the energy overhead, but also unleashes the interface

speed. For an in-memory layer with N_j inputs and N_k outputs, the number of linear operations consists of $N_j \times N_k$ complex multiplications and $(N_j - 1) \times N_k$ complex summations. Since each complex multiplication and addition corresponds to four and two real operations, respectively, the total linear operation number sums to $(6N_j\ N_k - 2N_k)$ while the number of nonlinear activations is $5N_k$, including one square activation (four real operations) and one nonlinear mapping. Accordingly, each FLARE layer performs $R = (6N_j\ N_k + 3N_k)$ operations in total.

On the other hand, the total energy consumption originates from two primary sources that draw energy from both electronic control circuits and the laser.

(1) Short-term memory. In each round of operations, the short-term memory is both written and read. In the memory writing process, with a measured short-term memory capacitance of 35 fF and an average charging voltage of 1.5 V, the energy cost of a nonlinear activation is given by $e_{st} = cV^2 = 78.6\ \mathrm{fJ/op}$. During memory readout, optical power is consumed to discharge the stored memory. The corresponding energy is expressed as $e_o = pt = (i/\eta)(cV/i) = cV/\eta$, where the η is the opto-electronic conversion efficiency, measured to be 0.5 A/W, yielding an optical energy cost of $e_o = 105\ \mathrm{fJ/OP}$.

(2) Long-term memory. Programming of the long-term memory requires digital-to-analog conversion, which constitutes the only DAC overhead in the computing process. In our experiments, two dedicated controller boards with 256 DAC channels were employed, consuming 11.4 W during the programming phase. Each programming cycle lasted 1 ms, after which the long-term memory could retain its state for up to seconds. Owing to the retention capability of the long-term memory, the programming lines can be disconnected after initialization, eliminating static power draw. As a result, the energy cost of long-term memory programming, when considering all inferences during the retention period, is $e_{lt} =$ 11.4 pJ per inference.

Given input and output sizes of N_j and N_k in a FLARE layer, based on the measurements of long-short term driving energy e_{st}, e_{lt}, and optical energy e_o, the layer-to-layer computing energy is $e = (N_j\ (e_{st} + e_o) + e_{lt})$, and the efficiency is calculated as e/R. Plugging in the input sizes of four layers as $N_1 = 38 \times 66$, $N_2 = 19 \times 33$, $N_3 = 9 \times 16$, and $N_4 = 4 \times 8$, the calibrated system performance is summarized in Supplementary Table 1, with per-layer energy efficiencies of 49.44 aJ/OP (attojoule per operation), 223.03 aJ/OP (femtojoule per operation), 1.16 fJ/OP, and 44.78 fJ/OP, and an overall system efficiency of 61.87 aJ/OP. Furthermore, the computational scale of this paradigm can be expanded by increasing the monolithic chip footprint and employing linearly scalable control circuitry.

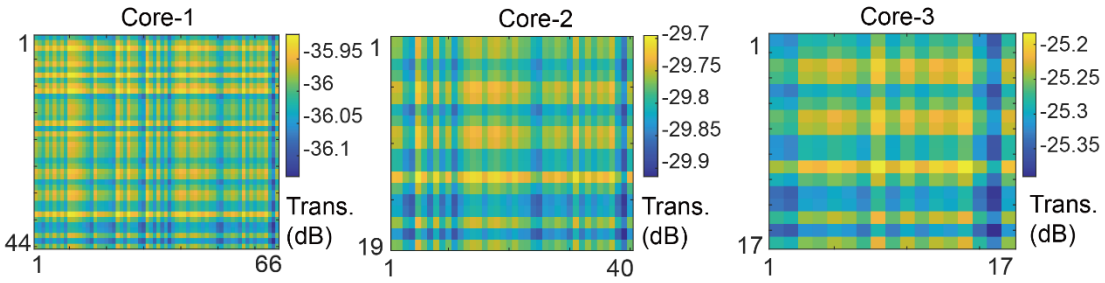

Fig. S2. Insertion loss of the FLARE neuron array circuit. From left to right, the losses correspond to Core-1, Core-2, and Core-3.

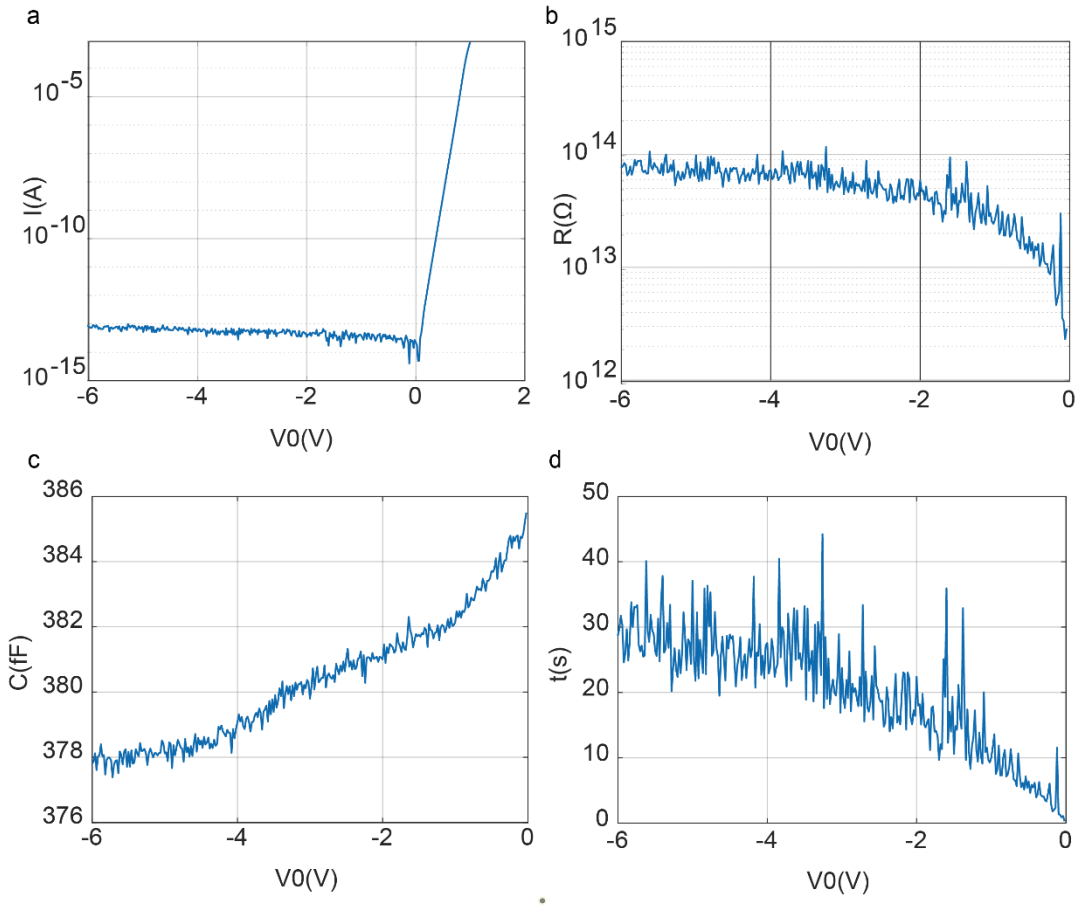

Fig. S3. The theoretical validation of long-term memory based on experimental measurement. **a,** Measured current–voltage (I–V) characteristics of the PN junction under reverse bias. **b,** The equivalent resistance of the reverse-biased junction. **c,** The measured capacitance of the long-term memory. **d,** The resistor-capacitor (RC) time constant.

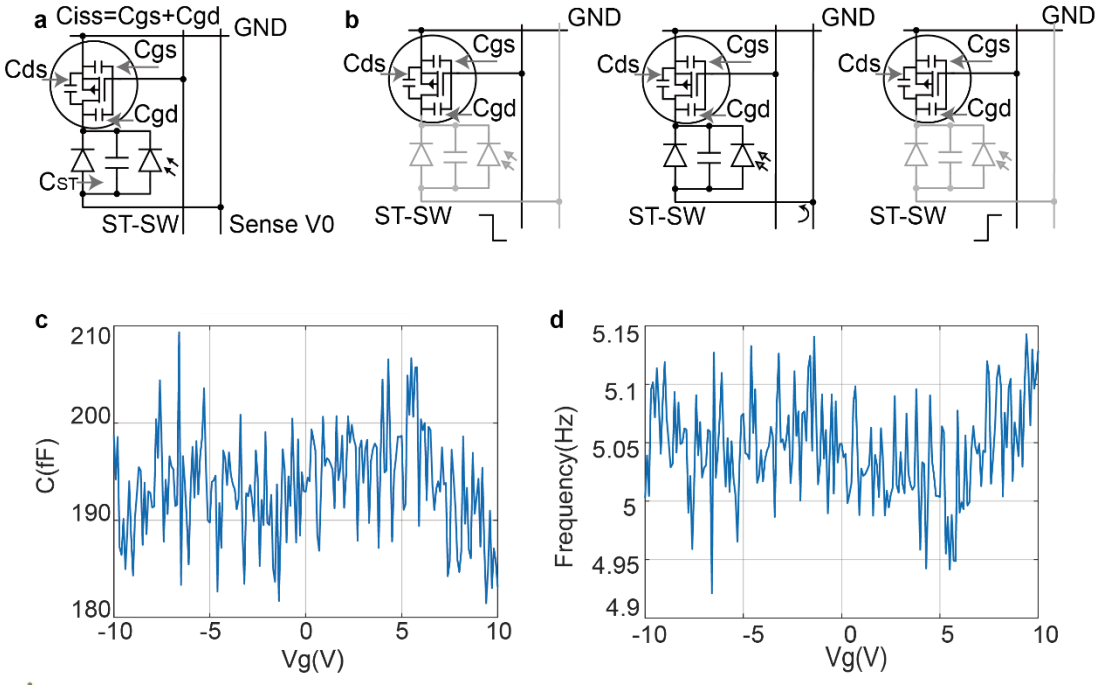

Fig. S4. Determination of the operating frequency of the ST-Mem FLARE neuron cell. **a,** Circuit schematic of the ST-Mem neuron. **b,** Switching and charging dynamics that govern the operating speed of the ST-Mem. **c,** Input capacitance of the MOSFET. **d,** Theoretically calculated operating frequency.

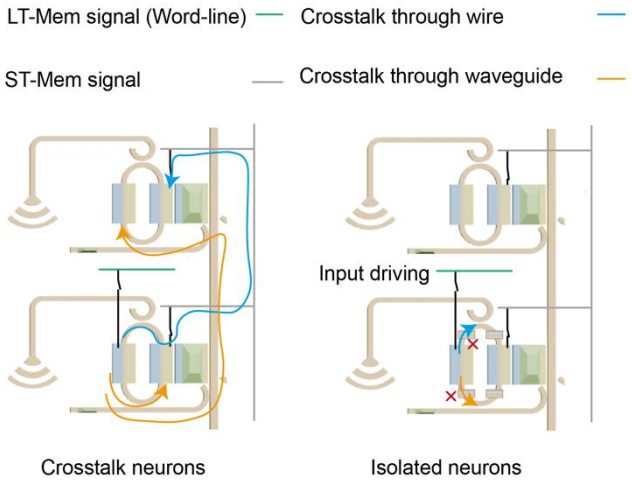

Fig. S5. Neuron isolation in optoelectronic circuits. Left: The driving signal leaks both inter- and intra-neuron through electrical interconnects and silicon waveguides, resulting in crosstalk and unintended perturbations of neighboring neurons. Right: With the proposed isolation architecture, signal leakage is effectively suppressed, enabling independent and reliable operation of individual neurons.

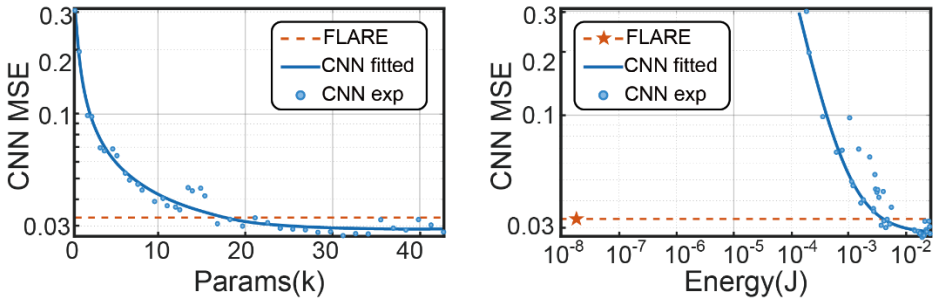

Fig. S6. Benchmarking the FLARE neural network against the electronic convolutional neural networks. The mean square error evolution with respect to the parameter number (Left panel) and energy (right panel).

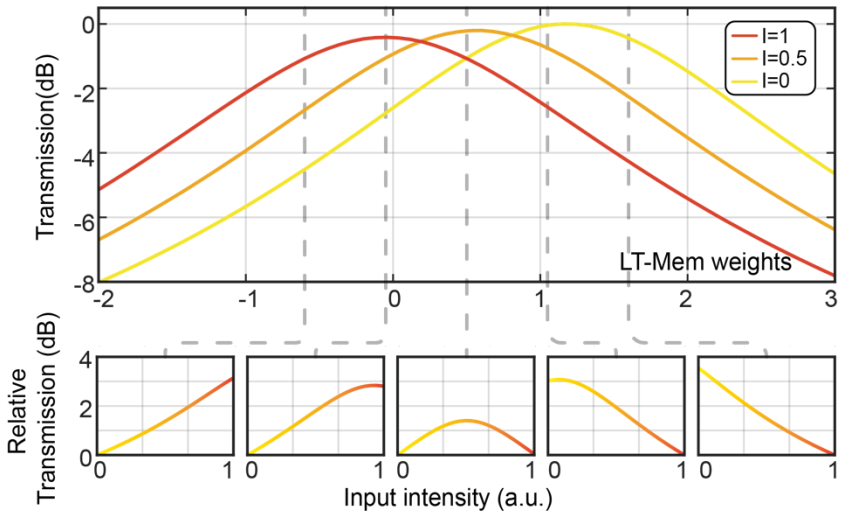

Fig. S7. The reconfigurability of the nonlinear transfer function. By programming different voltage weights to the LT-Mem unit and/or changing the probe wavelength, the operating point of each FLARE unit can be fully programmed, corresponding to different nonlinear activation functions. Top panel, the transmission curves at different incident intensities and LT-Mem weights (horizontal axis). Bottom panel, the transfer function when fixing the LT-Mem weights.

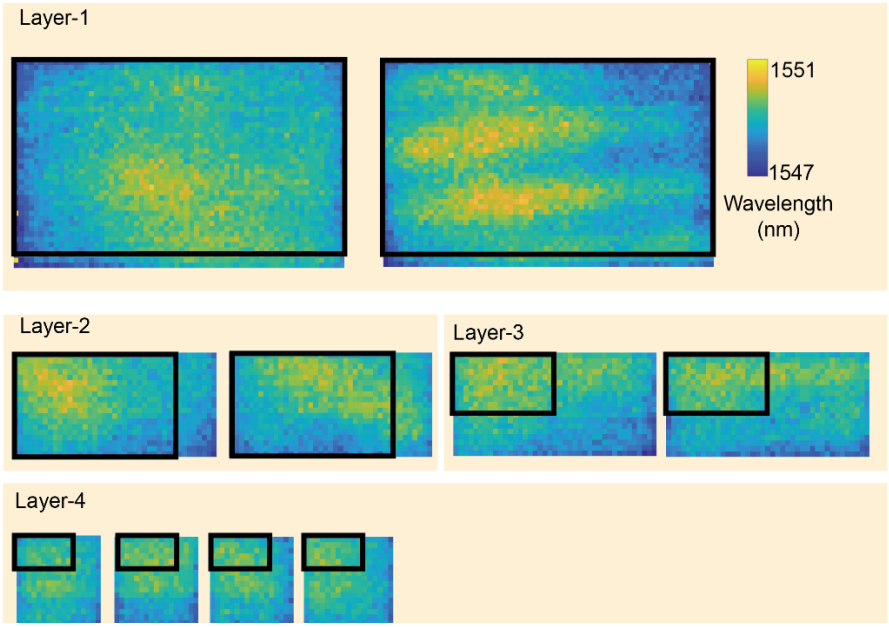

Fig. S8. The distribution of the resonant wavelength peak on the ten cores in the experiment. The black boxes delineate the selected functional sections of the chip.

Supplementary Table 1. Performance benchmarking.

		Layer-1	Layer-2	Layer-3	Layer-4	System
Number of operations (OP) per inference	Linear operations	18,867,684	1,082,880	55,168	760	20,006,492
	Nonlinear operations	6270	1440	320	20	33,430
	Summed operations	18,873,954	1,084,320	55,488	780	20,039,922
Energy of control circuits (Joule)	ST-Mem control per inference [b]	3.95×10^{-10}	9.88×10^{-11}	2.27×10^{-11}	1.01×10^{-11}	5.27×10^{-10}
	LT-Mem write per inference [a]	1.14×10^{-11}				
	LT-Mem write per operation [c]	6.04×10^{-19}	1.05×10^{-17}	2.07×10^{-16}	1.50×10^{-14}	5.70×10^{-19}
Energy of laser source (Joule) [b]		5.27×10^{-10}	1.32×10^{-10}	3.02×10^{-11}	1.34×10^{-11}	7.02×10^{-10}
System-level energy efficiency [d] (Joule per operation)		49.44×10^{-18}	223.03×10^{-18}	1.16×10^{-15}	44.78×10^{-15}	61.87×10^{-18}

a. LT-Mem units are managed with two shared controller boards equipped with low-
b. power DAC converters (DAC088S085) and an FPGA controller (XCKU060).
c. The ST-Mem controller and laser source collectively constitute the nonlinear activation energy, measuring as 1.84×10^{-13} Joule per operation on average.
d. Linear processing energy.
e. During one working cycle, power consumption of all system components: (1) LT-Mem controller board including FPGAs and DACs: 11.4 watts; (2) Single-channel ST-Mem driver: 532 mW; (3) Laser power: 702 mW.

Supplementary Table 2. Performance summary.

* Y, possesses the corresponding capability;

Techniques	Neuron Cascade	Ambient Light Sensing	Memory Scale per chip	Computing Throughput	Energy Efficiency	Packaging
Fully in-memory neuron	**Y**	**Y**	**7378**	**10.02 POPS**	**61.87 aJ/OP**	Wire bonding

**POPS, Peta (10^{18}) operations per second; TOPS, Tera (10^{15}) operations per second; aJ, attojoule (10^{-18} joule); fJ, femtojoule (10^{-15} joule); pJ, picojoule (10^{-12} joule); OP, operation.

***Throughput calculated based on one FLARE channel

Supplementary References

1 Zanetto, F. *et al.* Time-multiplexed control of programmable silicon photonic circuits enabled by monolithic CMOS electronics. *Laser & Photonics Reviews* **17**, 2300124 (2023).