







# Neuromorphic photonic computing with an electro-optic analog memory

Received: 28 February 2025

Accepted: 26 January 2026

Published online: 07 February 2026

 Check for updates

Sean Lam <sup>1</sup>✉, Ahmed Khaled <sup>2</sup>, Simon Bilodeau<sup>3</sup>, Bicky A. Marquez<sup>2</sup>, Paul R. Prucnal <sup>3</sup>, Lukas Chrostowski <sup>1</sup>, Bhavin J. Shastri <sup>2,3,4</sup>✉ & Sudip Shekhar <sup>1</sup>✉

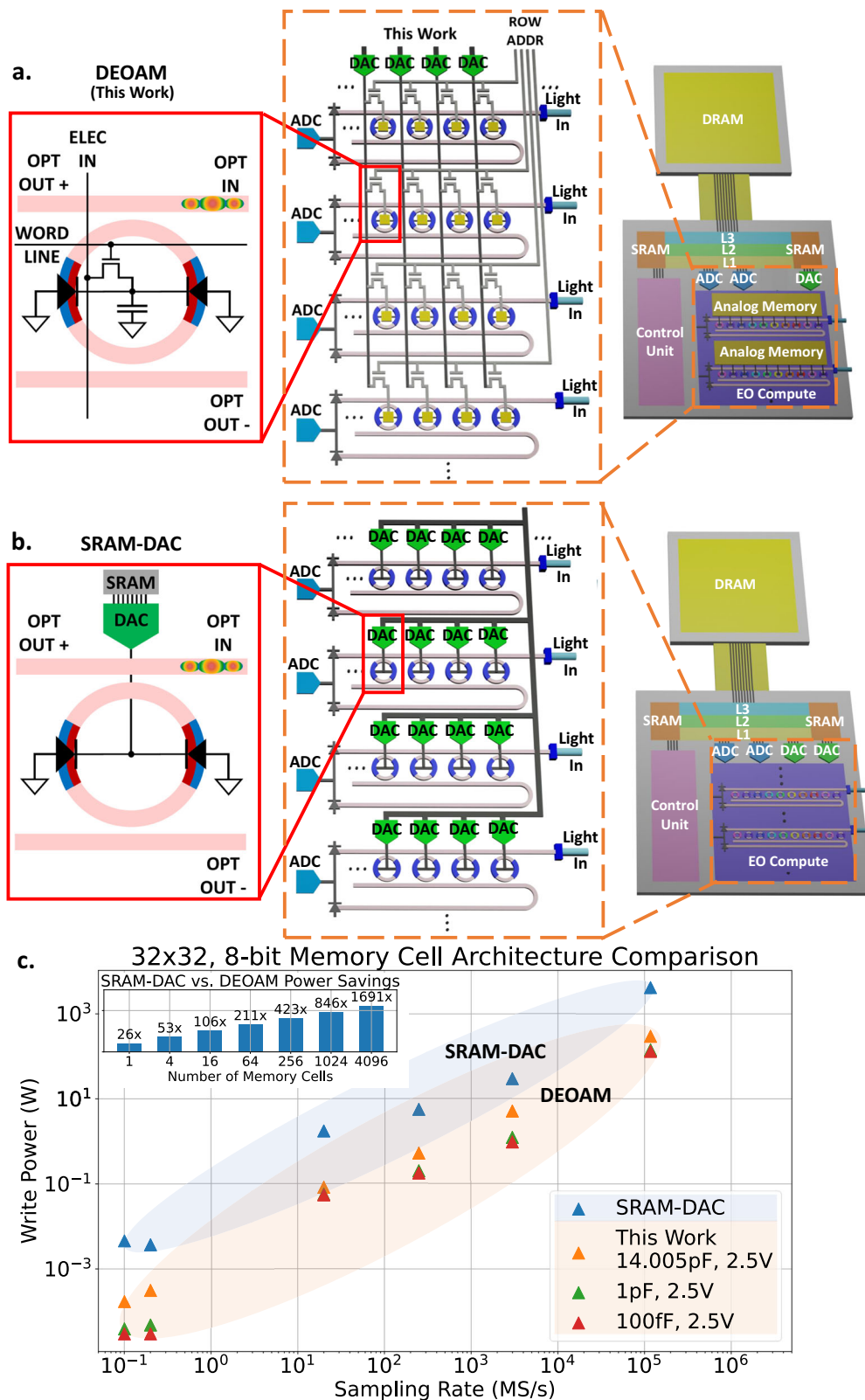
In neuromorphic photonic systems, device operations are typically governed by analog signals, necessitating digital-to-analog converters (DAC) and analog-to-digital converters (ADC). However, data movement between memory and these converters in conventional von Neumann architectures incur significant energy costs. We propose an analog electronic memory co-located with photonic computing units to eliminate repeated long-distance data movement. Here, we demonstrate a monolithically integrated neuromorphic photonic circuit with on-chip capacitive analog memory and evaluate its performance in machine learning for in situ training and inference using the MNIST dataset. Our analysis shows that integrating analog memory into a neuromorphic photonic architecture can achieve over  $26 \times$  power savings compared to conventional SRAM-DAC architectures. Furthermore, maintaining a minimum analog memory retention-to-network-latency ratio of 100 maintains  $>90\%$  inference accuracy, enabling leaky analog memories without substantial performance degradation. This approach reduces reliance on DACs, minimizes data movement, and offers a scalable pathway toward energy-efficient, high-speed neuromorphic photonic computing.

## Motivation

Artificial intelligence (AI) has profoundly influenced society and technology with advancements ranging from ChatGPT's human-like dialogs (<https://chat.openai.com/>) to AlphaFold's protein structure predictions<sup>1-3</sup>. Historically, AI's progression has paralleled advances in CMOS electronics, driven by Moore's law and Dennard scaling<sup>4</sup>. While Moore's law has slowed down, AI's performance demands have surged, with computational needs doubling every 2 months<sup>5</sup>. In parallel, neuromorphic engineering has emerged as a field aiming to align AI algorithms to hardware architectures that mirror their inherently distributed nature, with demonstrated applications in particle physics, nonlinear programming, and signal processing<sup>5</sup>. Among these innovations, neuromorphic photonics promise high bandwidth and low latency, providing a complementary opportunity to extend the

domain of AI<sup>5-8</sup>. The key advantages of neuromorphic photonic accelerators over purely electronic implementations include 1) massive parallelism using optical techniques, such as wavelength-division multiplexing (WDM), mode multiplexing (MDM), and time-domain multiplexing to carry multiple signals in a single interconnect and 2) lower energy-delay products in optical interconnects for multiply-accumulate (MAC) operations with potential speeds on the order of peta MACs per second per  $\text{mm}^2$  and energy efficiencies in the atto joule per MAC regime<sup>5,9-12</sup>. However, most photonic processors have been limited to stationary weights, which are trained offline on digital computers and remain unchanged during inference, thus restricting their application potential. One reason this limitation arises is because analog photonic processors require analog-to-digital (ADC) and digital-to-analog (DAC) converters (Fig. 1b), which become critical

<sup>1</sup>Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada. <sup>2</sup>Centre for Nanophotonics, Physics, Engineering Physics & Astronomy, Queen's University, Kingston, ON, Canada. <sup>3</sup>Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA. <sup>4</sup>Smith Engineering, Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada. ✉e-mail: [seanlm@student.ubc.ca](mailto:seanlm@student.ubc.ca); [shastri@ieeee.org](mailto:shastri@ieeee.org); [sudip@ece.ubc.ca](mailto:sudip@ece.ubc.ca)



**Fig. 1 | Motivation for electro-optic processors with dynamic electro-optic analog memory (DEOAM).** **a** Dynamic electro-optic analog memory (DEOAM) consists of a capacitor connected to a PN junction microring resonator (MRR). The capacitor holds data on the MRR, thereby enacting a weight in the optical domain. Using DEOAM means DACs can be shared amongst columns of analog memory devices, updating them row by row since the analog memory holds the signal on the MRR, which relaxes limitations on energy efficiency and bandwidth imposed by DACs. **b** Conventional electro-optic processors require a dedicated SRAM-DAC for

each PN junction MRR. The DAC needs to be constantly active to hold a signal on the MRR; therefore, implementing SRAM-DAC in a neuromorphic photonic processor requires a DAC for every MRR. **c** If  $n$  represents the number of MRRs in the row and column of a neuromorphic photonic processor, scaling the conventional SRAM-DAC implementation incurs an  $n^2$  DAC penalty, whereas DEOAM scales by  $n$ . Write power and sampling rate for  $32 \times 32$ , 8-bit memory cell architectures are compared using each memory device and DACs from literature<sup>22-30</sup>. As DEOAM scales, DEOAM saves more power compared to SRAM-DAC.

bottlenecks that affect bandwidth and energy efficiency as data transitions between digital memory and analog computation<sup>5,13–15</sup>. During training, backpropagation requires computations (gradient and error calculations) and memory accesses (weight updates and temporary data storage for calculations), meaning that any part of the pipeline that requires data converters to compute or access memory can be a bottleneck. This work focuses on data conversion bottlenecks from memory and showcases energy savings and minimized latency from co-locating analog memory with analog optical computational units.

Analog memory can reduce the number of DACs in neural network architectures (Fig. 1a). While optical analog memories (OAM) have been explored as in-memory computing devices, they currently face challenges, such as low yield and endurance, incompatibility with standard foundry processes, and complex fabrication processes<sup>16,17</sup>. Phase change materials (PCM) have been demonstrated in a photonic tensor core enabling  $10^{12}$  MACs per second for inference tasks<sup>9</sup>. However, the endurance of PCMs can vary depending on the modulation mechanism (electrical, electrothermal, or optical), raising concerns about their suitability for neuromorphic training applications that typically require volatile memory<sup>18</sup>. Optical memristors, which store memory by forming and dissolving filaments, offer minimal footprint, making them promising candidates for large scale integration. However, their limited endurance (~2000 write cycles) and challenges in compatibility with standard foundry processes remain key obstacles to broader adoption<sup>16,19</sup>. In contrast, analog electronic memories, such as dynamic random-access memory (DRAM), metal-oxide-metal capacitors, metal-insulator-metal capacitors, or metal-on-semiconductor capacitors, are foundry-compatible and included in conventional CMOS technologies. This compatibility makes integrating analog electronic memories into neuromorphic photonic hardware a more viable approach, potentially reducing the number of ADCs and DACs in applications, such as long-short term memory neural networks<sup>20</sup>.

Here, we propose a neuromorphic photonic processor with dynamic electro-optic analog memory (DEOAM, Fig. 1a). This processor leverages the monolithic integration of nanophotonic and CMOS devices on the same silicon substrate. The proposed architecture (Fig. 1a) merges the parallel information processing capabilities of photonics using wavelength-division multiplexing (WDM) with the low control complexity of crossbar arrays offered by electronics. As a proof-of-concept, we demonstrate neurosynaptic photonic weights, realized through a WDM array of tunable PN junction microring resonators (MRR) within a photonic weightbank and integrated with an array of capacitive DEOAM on the 90 nm GF9WG monolithic process. The MNIST dataset serves as a benchmark for training and inference and reveals tradeoffs in neural network design<sup>21</sup>. Our approach to integrating analog memory in neuromorphic photonic processors aims to create a flexible processor with high bandwidth, efficient training capabilities, and the ability to adapt through on-chip, online learning.

### Dynamic Electro-Optic Analog Memory (DEOAM)

DEOAM consists of a capacitor that holds charge on a reverse-biased PN junction MRR (Fig. 1a). Data is stored on the capacitor as an electrical signal then the PN junction MRR converts the signal to the optical domain for processing. DEOAM can be integrated into an array of MRRs, functioning as an electronic crossbar array and photonic weight banks (Fig. 1a) in a neuromorphic photonic processor. The conventional approach consists of several bits of static random access memory (SRAM) connected to a DAC that continuously drives a PN junction MRR (Fig. 1b).

In a neuromorphic photonic processor with  $n$  rows and  $n$  columns of MRRs acting as weights, DEOAM reduces the number of DACs, thereby reducing power consumption. Since the conventional SRAM-DAC implementation requires each MRR to be paired with a DAC,

scaling the implementation results in the number of DACs increasing by  $n^2$  (Fig. 1b). The power consumed by SRAM-DAC is calculated as:

$$P_{\text{SRAM-DAC}} = n^2(P_{\text{DAC-static}} + P_{\text{SRAM-static}}) + n(P_{\text{DAC-dynamic}} + P_{\text{SRAM-dynamic}}) \quad (1)$$

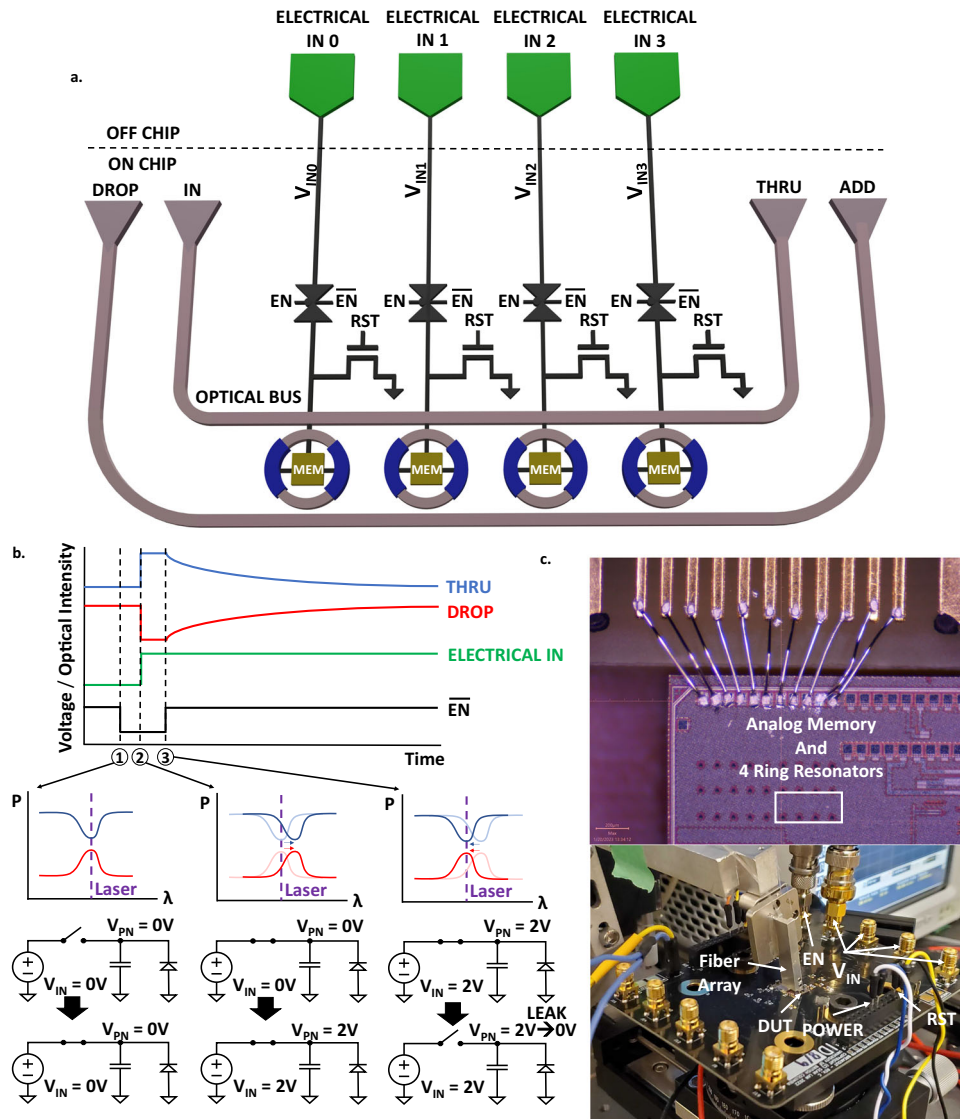
where  $P_{\text{SRAM-DAC}}$  is the total SRAM-DAC array power consumption in  $W$ , and  $P_{\text{SRAM}}$  and  $P_{\text{DAC}}$  are the SRAM and DAC power consumption in  $W$ , respectively, which includes static and dynamic power consumption and accounts for varying bit precisions of the DAC. All surveyed DACs are demonstrated, and the SRAMs associated with each DAC are surveyed to be in the same process node<sup>22–30</sup>.

In contrast, using DEOAM reduces the number of DACs below  $n^2$  (Fig. 1a) because data is held directly on the photonic device via analog memory, eliminating the constant processing typically needed by DACs. This concept is similar to the functionality in modern liquid crystal active-matrix displays<sup>31</sup>. The power consumed by DEOAM is calculated as:

$$P_{\text{DEOAM}} = n(P_{\text{DAC-static}} + P_{\text{DAC-dynamic}} + \alpha C f V^2) \quad (2)$$

where  $P_{\text{DEOAM}}$  is the total DEOAM array power consumption,  $\alpha$  is the activity factor (assumed to be 1),  $C$  is the load capacitance in  $F$ ,  $f$  is the switching frequency in  $Hz$  and is half the sampling rate, and  $V$  is the switching voltage. If we consider a memory element and DAC in both architectures, both architectures consume similar power when modulating the inputs (input to the SRAMs in the SRAM-DAC architecture and input to the DAC in the DEOAM architecture), but the main difference in power consumption is the SRAM to DAC power consumption in the SRAM-DAC architecture, the DAC to DEOAM power consumption in the DEOAM architecture, and the distribution of active DACs between the two architectures. In the SRAM-DAC architecture, power consumption begins when the SRAM changes state and ends when the DAC settles, meaning there is SRAM static and dynamic power and DAC static and dynamic power. In the DEOAM architecture, power consumption begins when the DAC changes state and ends when charge settles on the DEOAM, meaning there is DAC static and dynamic power and DEOAM dynamic power. Finally, there is the distribution of DACs from the architecture that determines the power consumption scaling. DACs are still required in the DEOAM architecture to load initial weights from non-volatile, digital memory before beginning inference or training and to refresh analog memory during inference. However, this approach not only simplifies control complexities by reducing the number of electrical lines needed for DAC operations but also, with longer retention times, diminishes the need for DACs to continuously drive the photonic devices, leading to further power savings as the processor scales (Fig. 1c, inset). As the DEOAM capacitance decreases, further power savings can be realized up to the limit of the junction capacitance of the PN junction (<50 fF<sup>32,33</sup>). Hence, integrating DEOAM and reducing the reliance on DACs paves the way for more efficient computing.

In neuromorphic processors, analog memory is typically situated close to or directly integrated with the computing unit to minimize data movement, thereby reducing energy consumption and latency<sup>34,35</sup>. Data movement in memory accesses can consume more than 10 times the energy of multiplication and addition operations<sup>36</sup>. Recent demonstrations of in-memory compute, such as Mythic's Analog Matrix Processor or NeuRRAM, achieve up to tens of TOPS/W in compute efficiency by co-locating the analog memory with the compute unit<sup>9,37–39</sup>. This highlights DEOAM's close integration with the MRR, which acts as the compute unit in the optical domain, thereby minimizing data movement and enhancing compute efficiency.



**Fig. 2 | The architecture and operation of the neuromorphic photonic prototype with integrated DEOAM.** **a** The analog memory circuit consists of a 1x4 MRR weight bank with each MRR connected to a capacitive analog memory. Transmission gates are used to enable/disable (EN) writing and switches are used to reset (RST) the analog memories. Laser light is injected into grating couplers through the IN port and is measured at the THRU and DROP ports. **b** The operation of the DEOAM circuit is shown. The PN junction MRR is modeled as a reversed biased diode, and the transmission gate is modeled as an ideal switch. At time 1, the laser

has been tuned to the MRR's resonance, and the transmission gate is connected such that the voltage source ( $V_{IN}$ ) can drive the analog memory ( $V_{PN}$ ). At time 2, voltage is driven onto the capacitor and MRR, causing the MRR's resonance to shift. At time 3, the transmission gate disconnects the voltage source from the capacitor, and the capacitor holds the charge on the MRR until it leaks away, returning the MRR resonance back to its original state. **c** The chip is wire bonded to a custom printed circuit board with thermal control and high-speed electrical lines. Details about the hardware operation and measurement setup are expanded in "Methods".

## Results

### Electro-optic characteristics

The neuromorphic photonic hardware with integrated DEOAM is shown in Fig. 2. The MRRs are PN-junction-doped modulators operating in depletion mode (reverse bias) and are connected to analog memory cells consisting of  $100 \times 100 \mu\text{m}$  metal-oxide-metal capacitors that use interdigitated fingers to increase capacitance. The MRRs have different ring perimeters starting with  $144.248 \mu\text{m}$  and  $50 \mu\text{m}$  PN junctions, increasing by  $50 \text{ nm}$  to avoid resonance collision. Key figures of merit for DEOAM are summarized in Table 1, and measurement results are shown in Fig. 3. Figure 2a, b show the DEOAM circuit and its operation. Weights are loaded as voltages on the capacitive memory, affecting the MRR's resonance. The capacitive memory holds the voltage on the MRR until the charge leaks away, returning the MRR to its resonance. Figure 2c shows the chip and measurement setup for static

optical and time domain measurements. Free spectral range, inter-channel spacing, Q-factor, and extinction ratio are crucial optical characteristics that determine the maximum number of MRRs and weights in a weight bank in a specific wavelength range. Therefore, if additional weights are needed in each row of a weight bank, solutions like adding more cores, reusing hardware, or employing time multiplexing may be needed for processing neural network data. In a neuromorphic photonic weight bank, light modulation is initiated at the input data modulators. The light then travels through the weight banks before being absorbed by photodetectors (PD) and amplified by transimpedance amplifiers (TIA). The group index helps characterize the group delay, which determines the overall compute time when combined with the delay from the PD and TIA. Since the delay from the PD and TIA depends on their design and technology, we report compute time based primarily on the group delay. State-of-the-art

**Table 1 | Experimental measurement results of the electro-optic analog memory circuit**

Parameter	Symbol	Min. Value	Nom. Value	Max. Value	Unit
Free Spectral Range	$FSR$	-	4.67	-	nm
Full Width Half Max.	$FWHM$	-	0.019	-	nm
Finesse	$F$	-	245.79	-	-
Inter-channel Spacing	$\Delta\lambda_0$	0.34	-	2.61	nm
Group Index	$n_g$	3.70	3.70	3.71	-
Q-Factor	$Q$	46050	49234	51484	-
Extinction Ratio	$ER$	15.6	-	19.9	dB
Wavelength Tuning Efficiency	-	-	6.22	-	pm/V
Modulation Efficiency	-	-	0.23	-	dB/V
Retention Time (10–90% Level)	$t_{ret}$	0.3330	0.5735	0.7957	ms
Retention Time (One Time Constant)	$\tau_{ret}$	0.2590	0.5527	0.8345	ms
Compute Time <sup>a</sup>	$\tau_{read}$	70.87	-	79.51	ps
Write Time - Rising Edge (10-90% Level)	$t_{rise}$	65.5	75.3	91.3	ns
Write Time - Falling Edge (10-90% Level)	$t_{fall}$	76.2	89.4	121.2	ns
Write Time - Rising Edge (One Time Constant)	$\tau_{rise}$	35.5	43.8	52.0	ns
Write Time - Falling Edge (One Time Constant)	$\tau_{fall}$	51.0	63.3	71.7	ns
Write Energy Consumption <sup>b</sup>	$\eta$	26.82	55.97	80.97	pJ/Write
Bit Precision	$N_b$	5.33	5.65	5.97	bits
Capacitance	$C_{MEM}$	-	14.005	-	pF

Write time and retention time measured with 0.5 dBm optical bus power in the circuit. Details regarding measurement setup and optical characteristics can be found in Methods and Supplementary Information.

<sup>a</sup> Estimated by the group delay through the circuit and resonance build-up in the MRRs. The summation of path delay ( $\tau_{path} = Ln_g/c$  where  $\tau_{path}$  is the path delay,  $L$  is the path length, and  $c$  is the speed of light) and MRR resonance build-up time ( $\tau_{MRR} = FRn_g/c$  where  $R$  is the ring radius) defines compute time<sup>73</sup>.

<sup>b</sup> Write energy consumption derived from maximum, minimum, and nominal retention time curves with nominal leakage curve at 0.5 dBm optical bus power in the circuit.

demonstrations of PDs and TIAs yield latencies ranging in tens of ps<sup>40–44</sup>. The wavelength and modulation efficiencies are also important for characterizing the tunability and range of weights for the neural network.

Key FoMs to characterize the capacitive analog memory are retention time, write time, read time, and power consumption. Since data written to DEOAM is immediately available in the optical domain, read time is effectively zero. Retention time is dependent on leakage, and leakage is dependent on optical power and photoresponse of the PN junction MRR<sup>45</sup>. Figure 3a, b reveals that the retention time of 0.8345 ms for one time constant can be extended if lower optical power is used since lower leakage is observed, but this involves a tradeoff between retention time and signal-to-noise ratio. Energy consumption, estimated to be 55.97 pJ/write, is linked to capacitor leakage and charge replenishment to maintain weight values. Figure 3c, d, and e shows that the write time is about 40–50 ns for one time constant, enabling MHz write speeds. From Fig. 3, the analog memory bit precision is about 5 bits and can be extracted from the time domain responses based on  $\log_2(\frac{\bar{\mu}_{max} - \bar{\mu}_{min}}{\sigma}) = N_b$  where  $\bar{\mu}_{max}$  and  $\bar{\mu}_{min}$  are the mean values at the maximum and minimum range of the analog memory, respectively,  $\sigma$  is the standard deviation, and  $N_b$  is the bit precision of the analog memory<sup>46</sup>.

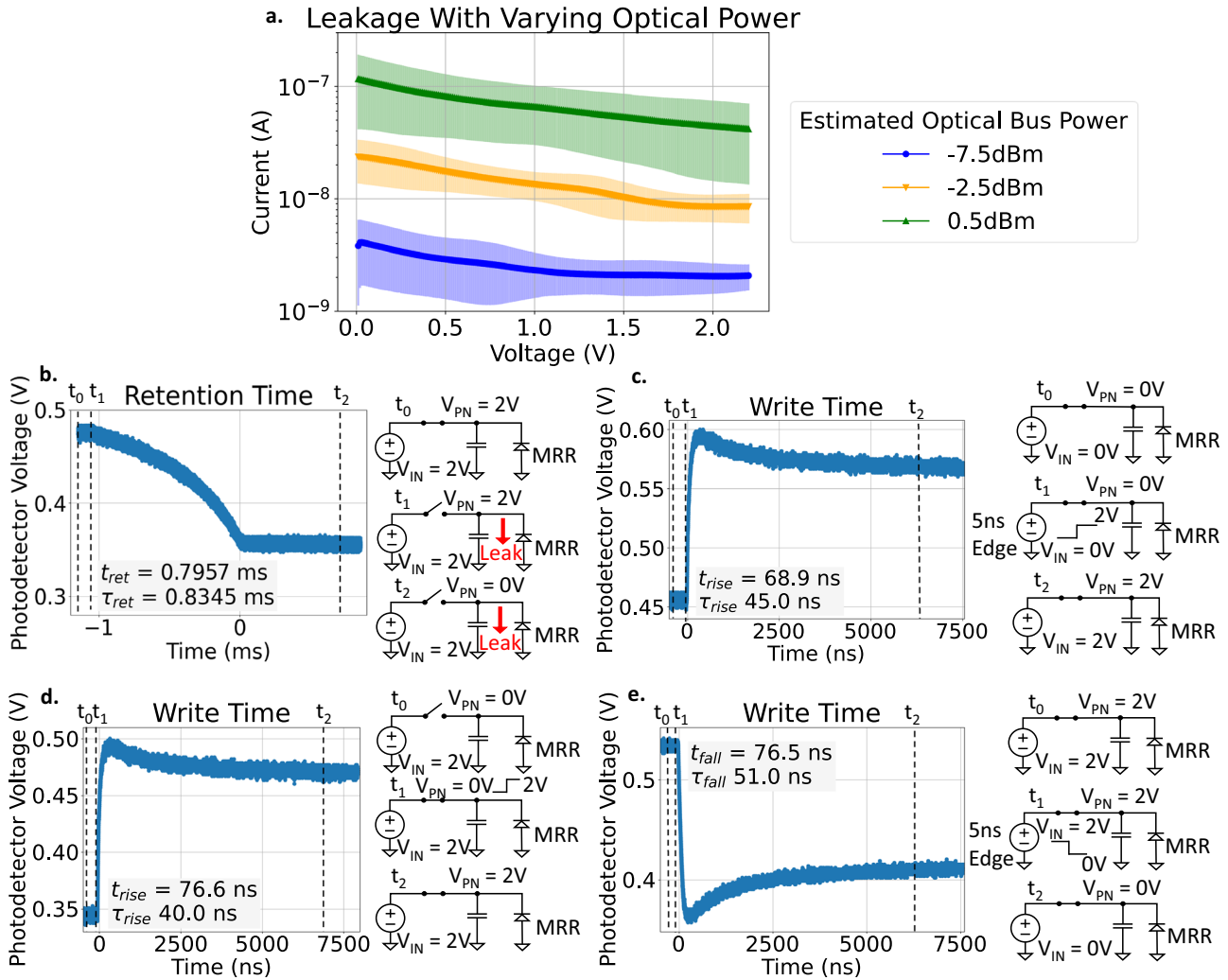
### Analysis of analog memory specifications

To verify different analog memory specifications within a neural network system, a weight bank architecture, depicted in Fig. 4, is emulated to assess analog memory performance during inference and training using the MNIST dataset<sup>21</sup>. The neural network architecture is configured as a three-layer model. The input layer processes 784 values corresponding to  $28 \times 28$  pixel image. This is followed by a hidden layer with 50 neurons and a ReLU activation function, and finally, an output layer with 10 neurons and a logarithmic softmax activation for digit classification. The network achieves over 95% validation and testing accuracy across all predicted digits within one

training epoch, as shown in Fig. 4b, which is comparable to the photonic neural network demonstrated in ref. 47. For a 5-bit resolution MRR-based weight bank matrix incorporating SOAs and MRRs with a finesse of 368 ( $\approx 55,000$  quality factor), the network size is limited to 108 rings (wavelength channels) per weight bank due to the MRRs' FSR and finesse. The number of rows (spatial channels) is limited to 60 due to the signal-to-noise ratio (SNR) constraints<sup>48,49</sup>. To ensure compatibility within a single core, the implementation uses 80 MRRs and 80 analog memories arranged in 50 rows of weight banks. Each weight bank requires two PDs and one TIA to perform the summation and signal amplification.

In Fig. 5, the weight bank array is characterized for inference accuracy by comparing inference accuracy using weights trained on a GPU with full floating point accuracy and weights trained on emulated hardware with non-idealities in the analog memory and hardware. These non-idealities include analog memory control bit precision, noise, latency, and retention time, which can be generalized to any analog memory technology. These non-idealities are examined individually to assess their impact on inference accuracy. Figure 5a reveals that to achieve inference accuracies above 80%, more than 3 control bits are required for analog memory when weights are trimmed only in inference. The minimum control bit requirement increases to over 5 control bits when weights are adjusted in training and inference because lower control bit precision during training results in less accurate gradient calculations and weights, leading to reduced inference accuracy.

Figure 5b, c analyzes the impact of noise sources on inference accuracy after training without noise and with noise, respectively. The main noise sources that have been taken into account come from the laser, SOA, PD, and TIA. Results indicate that inference accuracy degrades with more than 12.5% optical power noise injected into the weight banks. Furthermore, networks trained with noise are more resilient to PD noise current during inference than networks trained without noise. Specifically, the network trained with noise maintains



**Fig. 3 | Electro-optic characteristics of the DEOAM. a** Leakage shows a dependence on optical power incident on the doped MRR. Error bands represent the standard deviation in the leakage. **b** Retention time from leakage is 0.835 ms for one time constant ( $\tau_{ret}$ ) at 0.5 dBm optical bus power. The decay shows a nonlinear response from the ring's Lorentzian response. **c** With the transmission gate on, 2 V is written to analog memory using a 5 ns edge rate, and the resulting write time is

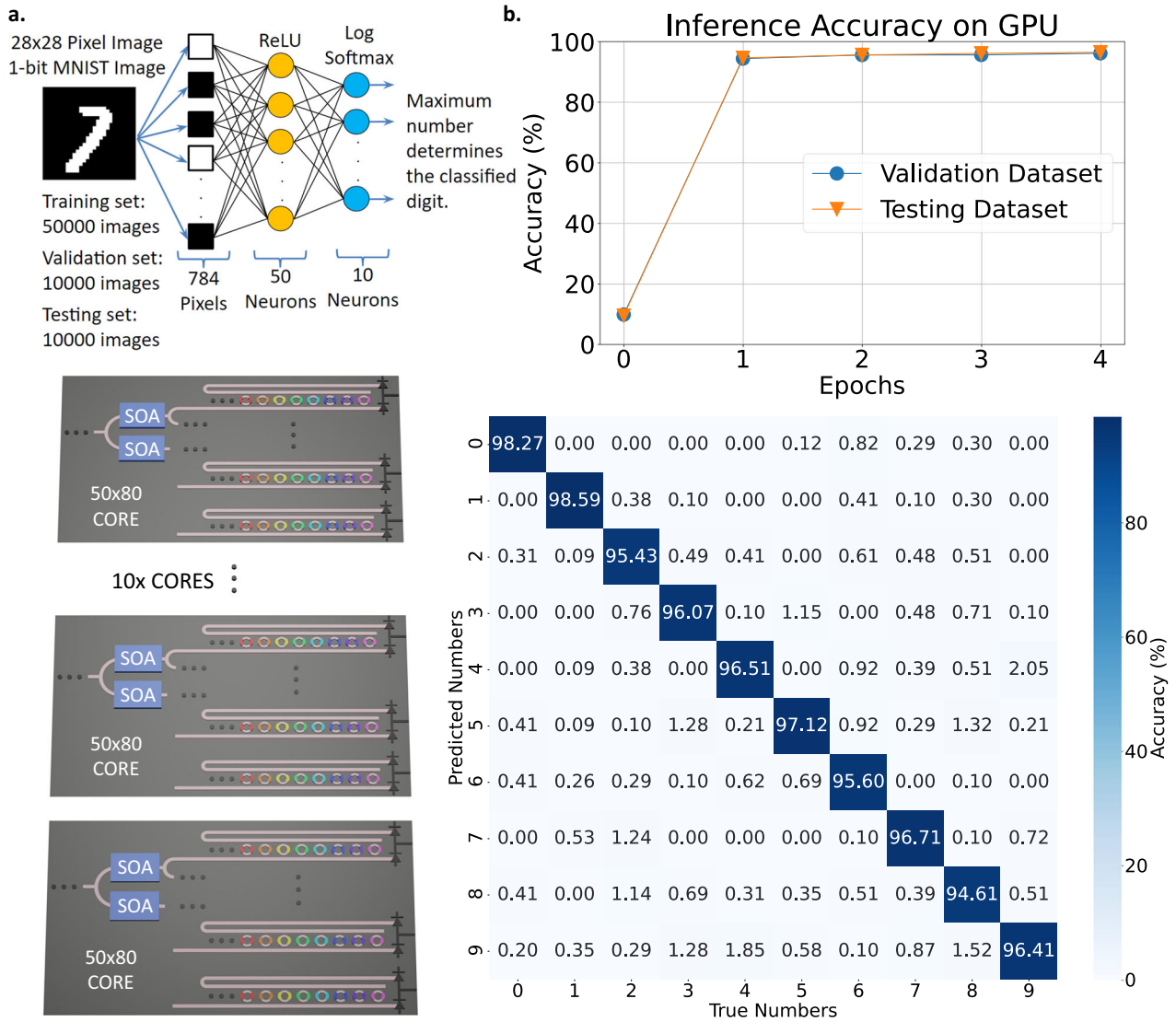
45.0 ns for one time constant ( $\tau_{rise}$ ). **d** With a source voltage of 2 V on  $V_{IN}$  and toggling the transmission gate from off to on, write time is 40.0 ns ( $\tau_{rise}$ ). **e** With the transmission gate on and writing 0 V to analog memory at an edge rate of 5 ns, write time is 51.0 ns for one time constant ( $\tau_{fall}$ ). Further discussions on the electro-optic results are expanded in Supplementary Information.

high inference accuracy until the PD noise current reaches about 2 mA RMS, whereas the network trained without noise only maintains high inference accuracy until about 0.1 mA RMS of PD noise current.

As weights decay due to the analog memory's leakage, the memory's retention time constant and the network latency determine the number of images that can be classified correctly or the number of useful computations before the weights become unusable. Network latency represents the time for an image to pass through the network. Within a retention time constant,  $k$  number of images can be passed through the network before weights are unusable, incurring  $k$  times the network latency. This indicates that the ratio of retention time constant to network latency is crucial for a network's inference performance.

Figure 5d examines the impact of analog memory retention time constant (weight decay) and network latency on inference accuracy when it degrades by 10%. The weight decay is modeled as an exponential decay based on the ratio of retention time constant to network latency. These experiments compare scenarios where weights are trained with and without considering retention time and network latency. The findings suggest that training with retention time constant

and network latency relaxes the requirements for analog memory retention time, enabling the use of memories with shorter retention time. A ratio of retention time constant to network latency greater than 100 prevents inference accuracy from decaying by 10%. This implies that even with an analog memory retention time constant of 100  $\mu$ s, if the network latency is less than 1  $\mu$ s, the network can maintain high inference accuracy. The ratio of retention time constant to network latency can be scaled to any analog memory technology, giving either a minimum retention time constant for a given network latency or a maximum network latency for a given retention time constant. Although analog memory leakage may limit inference accuracy, controlled leakage can be employed as a regularization technique to avoid overfitting, enabling broader inference applicability and meta-learning<sup>50,51</sup>. Furthermore, retention time constant and network latency are varied to evaluate training accuracy with different batch sizes, including 16, 32, and 64. The results show that larger batch sizes require longer retention times to achieve similar training accuracy as smaller batch sizes. With smaller batch sizes, weights are updated more frequently, so the analog memory does not need to retain the data for a long time and can operate with shorter retention times.



**Fig. 4 | The neural network emulation architecture and initial training.** **a** The feedforward neural network is trained using the MNIST dataset with 50,000, 1-bit, 28 × 28 pixel images using a batch size of 64, then validated and tested using 10,000 images<sup>21</sup>. The neural network is three layers with the input layer supporting 784 values for the 28 × 28 pixel images, the hidden layer supporting 50 neurons using a ReLU activation function, and the output layer supporting 10 neurons using a logarithmic softmax activation function to classify as a number. The neural

network can map to the proposed neuromorphic photonic hardware that consists of 10 photonic cores with 50 rows of weight banks and 80 MRRs in each weight bank. Each core uses 80 wavelengths and 50 semiconductor optical amplifiers (SOA) to compensate for the splitting loss. **b** The neural network achieves more than 95% inference accuracy after one epoch for all numbers. The specifics of the neural network architecture and modeling are detailed in “Methods”.

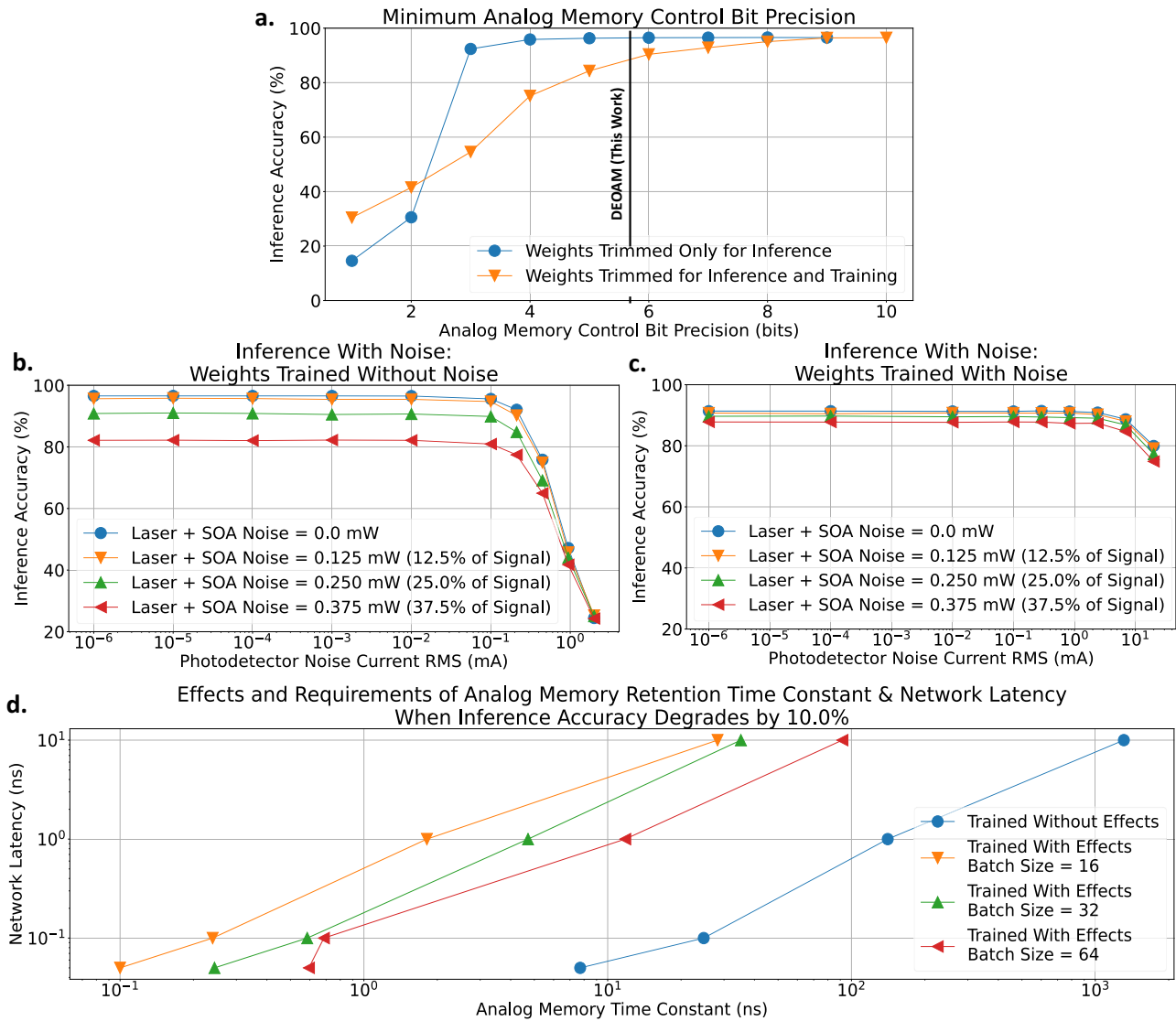
Once weights become unusable due to leakage, dead time is consumed to refresh the weight, which consists of the time needed to update the DAC inputs, analog memory write time, and DAC settling time. For DEOAM, the write time is about 65 ns in the fastest case. Surveyed DACs, in similar process nodes to DEOAM, can achieve sample rates up to 3 GS/s (or 333 ps per sample), and reported SRAM propagation delays are in the 10s of picoseconds, meaning that the time needed to update the DAC inputs and DAC settling time is negligible compared to the DEOAM write time<sup>24,25</sup>. Even for the fastest write time among surveyed analog memory technologies (500 ps for PCMs), the analog memory write time is similar to DAC sample rates<sup>52</sup>. Therefore, the dead time can be approximated by the analog memory’s write time. When examining the ratio of retention time to write time for analog memories, this ratio (10,000 for DEOAM) is a couple of orders of magnitude greater than the retention time-to-network-latency ratio (100) required for high inference accuracy. This means

the dead-time for refreshing can be much shorter than network latency. Therefore, refreshes can occur after each image classification, thereby maintaining accurate weights for high inference accuracy without adding significant latency, but at the cost of refresh power.

### Discussion

Training neuromorphic photonic processors presents several challenges related to energy efficiency, time, and execution. Analog-to-digital and digital-to-analog conversions, along with memory access, often become energy and latency bottlenecks. Additionally, since modeling hardware and environmental fluctuations are challenging, relying on offline-trained weights can degrade inference accuracy<sup>13,15,53</sup>.

To address these challenges, co-locating analog memory with the compute unit can alleviate energy consumption and latency by reducing data movement and the need for ADCs and DACs. Recurrent neural networks, in particular, can benefit from analog memory



**Fig. 5 | Neural network emulation results with and without non-idealities from the analog memory and system, such as control bit precision, noise, leakage, and latency.** **a** Minimum analog memory control bit precision to achieve more than 95% inference accuracy is 4 bits for weights trimmed only in inference and 8 bits for weights trimmed in inference and training. Inference accuracy degrades with increasing laser, SOA, PD, and TIA noise, but inference accuracy degrades **b** earlier when weights are trained without noise and **c** later when weights are trained with noise. **d** The sweep of network latency and retention time characterizes inference accuracy for weights trained with and without leaky analog memory and network

latency. Network latency and analog memory retention time constant at the point when inference accuracy degrades by 10% are plotted. When weights are trained with the effects of network latency and memory retention time, the network compensates for these effects and allows for a lower analog memory retention time constant requirement. Batch sizes of 64, 32, and 16 reveal that large batch sizes require longer retention times since weights are updated less frequently. Weights are updated more frequently in smaller batch sizes. The Supplementary Information expands on results in **(d)**.

integration since retaining data in the analog domain reduces analog-to-digital and digital-to-analog conversions<sup>20</sup>. Furthermore, analog memory can enhance the effectiveness of other training algorithms, such as direct-feedback alignment and in-situ training, by improving training time and on-chip computations<sup>15,54</sup>. A potential outcome from integrating analog memory in neuromorphic photonic processors is on-chip, online training, where the processor can adapt to new data coming in, be more resilient to noise and hardware flaws, and consume less energy<sup>53</sup>. Architecturally, incorporating analog memory into neuromorphic photonic processors facilitates system-level improvements, leading to improved energy efficiency, shorter training time, and a more adaptive neural network.

However, scaling the neuromorphic photonic processor with analog memory presents challenges in analog memory and MRR

control. As the network scales, reasonable sizes of the network can be up to 108 rings in a weight bank (limited by the ring's FSR and finesse) and up to a few hundred rows of weight banks (limited by SNR)<sup>48,49</sup>. The size of the network can also be limited to the analog memory's retention time, refresh time, and inference time (network latency) where a retention time must be allocated to refresh time for all rows of the network and inference time to ensure the network can operate. For DEOAM, the retention time to write time ratio is about 10,000, and inference times (picosecond to nanosecond range) are negligible compared to the write time. Therefore, if SNR is not limiting the number of rows in the network, the next limitation is the refresh rate, which is limited to several thousand rows. Depending on the analog memory requirements, specialized circuitry may be necessary for writing to or refreshing the memory. As the system scales, strategies

like sharing drive lines (as illustrated in Fig. 1a) where DACs are shared along columns, can be essential in reducing power consumption. Accurate resonance control in MRRs is vital for maintaining precise optical weight representation. As the processor scales, implementing thermal control and multi-MRR control algorithms becomes increasingly critical for system stability<sup>33,46,55–58</sup>.

While analog memory can provide system-level enhancements, analog memory specifications must be examined carefully to understand their limitations. Analog memory endurance is particularly critical, as it must endure numerous writes during training. Selecting training algorithms to minimize writes can enhance endurance, but choosing the right analog memory technology for the application is crucial for determining network endurance. Despite DEOAM's leakage, high inference accuracy can still be achieved if the ratio of retention time-to-network latency is 100 or higher. This and other charge-based or leaky analog memories can remain effective and maintain high inference accuracy despite inherent leakage.

The modulator is another area for enhancement. Employing PN junctions with different doping levels and junction geometries, such as interleaved, U-shaped, or L-shaped designs, can improve tuning efficiency and speed<sup>56,59,60</sup>. However, as indicated in Fig. 3a, the PN junction MRR exhibits leakage that is dependent on the incident optical power, suggesting that the photoresponsivity of the PN junction can limit analog memory performance. Reducing this leakage may involve removing contaminants and impurities, such as layering poly-silicon on the substrate or optimizing convection annealing<sup>61,62</sup>. Alternatively, other modulators, such as MOS capacitor modulators, can be explored to potentially reduce leakage<sup>63</sup>. Pockels effect modulators might offer advantages over PN junctions, as they exhibit minimal leakage (less than 100 nA) and better tuning efficiency<sup>64–66</sup>. Further development in foundry processes is needed to better integrate these modulators into electro-optic platforms<sup>67</sup>.

We demonstrated a proof-of-concept neuromorphic photonic circuit with DEOAM and verified its capabilities in a neuromorphic system. DEOAM reduces DACs in a neuromorphic photonic processor from  $n^2$  to  $n$ , thereby alleviating bottlenecks in bandwidth and energy consumption. By integrating analog memory on chip, neuromorphic photonic processors have the potential to achieve fast and efficient training. Since the training process occurs on chip, the neural network can readily adapt to environmental fluctuations and noise by incorporating them during training. Moreover, analog memory with on-chip backpropagation hardware can enable online training, allowing weights to be updated after each new input data. This opens up possibilities for various applications requiring recurrent events, event-based decision-making, or adaptability, such as time series forecasting. As a result, the next generation of AI processors, equipped with integrated analog memory, can help sustain the continued growth of AI.

## Methods

### Hardware

In Fig. 2a, the silicon photonics electronic chip, fabricated in the 90 nm GF9WG monolithic process, consists of a  $1 \times 4$  MRR circuit with C-band grating couplers to couple light to the chip and electronics to control when to load or reset voltages. The electronics consist of transmission gates and reset (RST) transistors. The transmission gates are controlled by an active low enable (EN) signal, which connects (EN = 0) or disconnects (EN = 1) an external source measure unit (SMU) to the analog memory cells. The RST transistors reset the voltage on the analog memory cells before operation. From Fig. 2c, the silicon photonic electronic chip is thermally glued and wirebonded to a custom printed circuit board (PCB) that breaks out connections to measurement equipment. The PCB has 50  $\Omega$  transmission lines that breakout to subminiature version A (SMA) connectors for high speed signals and 2.54 mm pitch headers for power and low-speed signals. A 10k negative temperature coefficient (NTC) thermistor and peltier module are

connected to the LDC501 for thermal control. An 8-degree, 8-channel fiber array is used to couple light to the chip.

Figure 2b shows the operation of the circuit. Initially, the laser is aligned with the MRR resonance while the MRR is voltage-biased at zero voltage. At time 1, the EN signal is active, allowing voltages from the SMU to propagate to the analog memory. At time 2, the SMU drives the capacitor with an analog voltage, which also drives the PN junction MRR. At time 3, when the EN signal is inactive, the capacitor holds the analog voltage on the PN junction MRR until the charge leaks away. As a result, the optical signals on the through (THRU) and drop (DROP) ports of the MRR circuit track the voltage on the capacitor and PN junction MRR.

### Measurement setup

From Fig. 2a, the optical characteristics are measured using a C-band tunable laser (HP81682A) and an array of photodetectors (N7744A) while the MRRs are voltage-biased using an SMU (Keithley 2604B). Light is coupled through an 8-degree, 8-channel fiber array to C-band grating couplers on chip, and the SMU sweeps voltage to observe MRR resonance changes with varying voltage biases.

For the retention time measurement setup, the HMP4040 power supply unit powers the electronics, the LDC501 thermoelectric cooler (TEC) regulates the chip temperature, and the HP81682A tunable C-band laser provides optical power to the circuit for optical processing. To drive signals, the Keithley 2604B SMU is used to drive the analog memory and PN junction MRR and the Agilent 33250A arbitrary waveform generator (AWG) drives the EN signal with a 5 ns edge rate step to quickly connect and disconnect the SMU from the analog memory and PN junction MRR. The DROP port is connected to a PD on the N7744A to tune the laser wavelength to the MRR resonance, and the THRU port is connected to the high-speed, amplified PDA255 and DSO81304A oscilloscope to measure the retention time of the circuit. The retention time measurement sequence is shown in Fig. 3b. At  $t_0$ , EN is active, and the SMU sets the desired voltage on the analog memory and PN junction MRR. At  $t_1$ , EN becomes inactive, and the analog memory retains the data on the PN junction MRR. After  $t_1$ , charge leaks away from the analog memory, which returns the optical response on the THRU and DROP ports back to its original state (back to the MRR's resonance). At  $t_2$ , all charge is depleted on the analog memory and the optical response returns to the MRR's resonance. The leakage measurement setup is the same as the retention time setup, except that the EN signal is constantly active and the SMU is sweeping the voltage from 0 to 2 V while measuring leakage current through the circuit.

The write time measurement setup is similar to the retention time and leakage measurement setup, except that the AWG is driving the analog memory and PN junction MRR while the SMU is holding the EN signal active. The write time measurement sequence is shown in Fig. 3c, e. The AWG sends a pulse train with 5 ns edge rates to drive the analog memory and PN junction MRR, which modulates the optical response on the THRU and DROP port. The optical response is detected using the THRU port by the PD255 and measured by the DSO81304A oscilloscope. The write time measurement sequence shown in Fig. 3d differs from Fig. 3c, e because the rather than driving the analog memory and PN junction with the 5 ns edge rate pulse train, the transmission gate is driven with a pulse train that connects and disconnects the voltage source from the analog memory and PN junction. The rise time, fall time, and one time constant values are extracted from the oscilloscope time domain response.

### Neural network architecture

To map the neural network onto hardware, the input data consists of  $28 \times 28$  1-bit images (784 pixels), which are mapped to 80 input data modulators per core across 10 cores, for a total of 800 modulators, each handling a single pixel (with some leftover). Each pixel (input

data point) is multiplied by 50 different weights to connect to the hidden layer. Since each core contains 50 rows of weight banks, there are 50 MRRs or weights associated with each pixel. Once an input pixel is multiplied by a weight, summation occurs at the PD within each core and then across cores to obtain the 50 neurons in the hidden layer. In each core, 80 pixels are processed (multiplied by their weights) and summed at the PD in each row of weight banks. To process all 784 pixels for the hidden layer, the summation from a given row in one core is combined with the corresponding row in the other nine cores. This yields the total weighted sum for that row, effectively multiplying 50 weights by each of the 784 pixels to produce the inputs to the 50 hidden neurons. A ReLU activation function is assumed to be implemented in analog electronics. The resulting 50 neurons generate 50 intermediate signals, which are passed to the input data modulators for processing in the output layer. In this stage, each signal is multiplied by 10 weights (10 MRRs), allowing the entire multiplication and summation to fit within a single core. After summation, a logarithmic softmax activation (also assumed to be implemented in analog electronics) classifies the signals to predict the corresponding digit.

### Noise

Laser noise represents input data noise and can be related to relative intensity noise (RIN) from the laser as shown in the following equations<sup>68</sup>:

$$x_i = \bar{x}_i(1 + dx_i) = \bar{P}_i(1 + dP_i) = \bar{P}_i(1 + \mathcal{N}(0, 10^{RIN/10} \cdot f)) \quad (3)$$

where  $x_i$  is the  $i$ th input data,  $\bar{x}_i$  is the average value of the  $i$ th input data,  $dx_i$  is the  $i$ th random noise for the  $i$ th input data,  $\bar{P}_i$  is the  $i$ th average laser power,  $dP_i$  is the  $i$ th random noise in the laser power,  $\mathcal{N}$  is the normal distribution function,  $RIN$  is the  $i$ th relative intensity noise in dB/Hz, and  $f$  is the integration bandwidth in Hz. As laser light splits to reach each weight bank, splitting loss is incurred. SOAs can amplify the signal to compensate for splitting loss but adds noise to the input data. As a result, the input data to each weight can be described by:

$$x'_i = x_i + n_{SOA_j} \quad (4)$$

where  $x'_i$  is the  $i$ th input data to the  $i$ th weight that includes SOA and laser noise, and  $n_{SOA_j}$  is the  $j$ th row SOA noise. As the input data is multiplied by weights, insertion loss from MRRs degrades the input data signal. Weight bank insertion loss ( $\beta$ ) is modeled by using an insertion loss of 0.125 dB/MRR, giving a total insertion loss of 10 dB in each weight bank. At the input of each weight bank, each wavelength has 1 mW of optical power and the noise ranges from 0% to 37.5% of 1 mW to mimic noise from the laser and SOAs. Furthermore, noise is applied to the data after the MAC operation and before the activation function to mimic the input-referred noise current on the PD and TIA. The PD and TIA integrated noise current ranges from 1 nA to 2 mA based on current PD and TIA technology<sup>40–44,69–72</sup>. The following equation summarizes the noise sources and the neural network equation:

$$y_j = \sigma_j(\beta \mathbf{x}' \cdot \mathbf{w}_j + n_j) \quad (5)$$

where  $y_j$  is the  $j$ th output data in a layer,  $\beta$  is the weight bank insertion loss,  $\sigma_j$  is the  $j$ th activation function in a layer,  $\mathbf{x}'$  is the input data vector to the weight banks that includes SOA and laser noise ( $\mathbf{x}' = \langle x'_0, x'_1, \dots, x'_i \rangle$ ),  $\mathbf{w}_j$  is the  $j$ th vector of weights in a layer, and  $n_j$  is the  $j$ th input referred noise term. The dot product of  $\mathbf{x}'$  and  $\mathbf{w}_j$  represents the photocurrent from the PD detecting laser wavelengths weighted by MRRs.

### Weight decay and network latency

For each weight, weight decay is modeled by using a ratio of retention time to network latency in an exponential decay model to determine inference accuracy from leaky weights:

$$w(k) = w_0 e^{-k \frac{t_{\text{latency}}}{\tau_{\text{ret}}}} \quad (6)$$

where  $w(k)$  is the weight value for the  $k$ th image,  $w_0$  is the initial weight,  $t_{\text{latency}}$  is the network latency or compute time,  $\tau_{\text{ret}}$  is the retention time constant. In inference, network latency is determined by group delay from photons propagating through the optical circuit and driver, modulator, PD and TIA bandwidth, whereas training has additional latency from analog electronics used to update weights. Since the bandwidth of the PD and TIA and analog electronics can vary depending on design and technology nodes, the range of network latencies is assumed to be determined by the PD and TIA bandwidth, which translates to latencies ranging from 50 ps to 10 ns<sup>40–44,69–71</sup>.

### Data availability

Methods, source data, experimental setup, and additional analyses and results are provided in the Supplementary Information and are available from the corresponding authors on request. The measurement data and source code in this study have been deposited in the Figshare database under the accession code <https://doi.org/10.6084/m9.figshare.30176863>.

### References

- AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).
- Merolla, P. A. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
- Bannon, P., Venkataramanan, G., Sarma, D. D. & Talpes, E. Computer and redundancy solution for the full self-driving computer. In: *Proc. IEEE Hot Chips 31 Symposium (HCS)* 1–22 (IEEE, 2019).
- Mehonic, A. & Kenyon, A. J. Brain-inspired computing needs a master plan. *Nature* **604**, 255–260 (2022).
- Shastri, B. J. et al. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photonics* **15**, 102–114 (2021).
- Ferreira de Lima, T. et al. Primer on silicon neuromorphic photonic processors: architecture and compiler. *Nanophotonics* **9**, 4055–4073 (2020).
- de Lima, T. F. et al. Machine learning with neuromorphic photonics. *J. Lightwave Technol.* **37**, 1515–1534 (2019).
- Tait, A. N. et al. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **7**, 7430 (2017).
- Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
- Khaled, A., Aadhi, A., Huang, C., Tait, A. N. & Shastri, B. J. Fully integrated hybrid multimode-multiwavelength photonic processor with picosecond latency. *Nat. Commun.* **17**, 28 (2024).
- Lin, Z. et al. 120 GOPS Photonic tensor core in thin-film lithium niobate for inference and in situ training. *Nat. Commun.* **15**, 9081 (2024).
- Ou, S. et al. Hypermultiplexed integrated photonics-based optical tensor processor. *Sci. Adv.* **11**, eadu0228 (2025).
- Sun, Z. et al. A full spectrum of computing-in-memory technologies. *Nat. Electron.* **6**, 823–835 (2023).
- Antonik, P. Online training of a photonic reservoir computer. In Antonik, P. (ed.) *Application of FPGA to Real-Time Machine Learning: Hardware Reservoir Computers and Software Image Processing*, Springer Theses, 39–62 (Springer International Publishing, 2018).

15. Filipovich, M. J. et al. Silicon photonic architecture for training deep neural networks with direct feedback alignment. *Optica* **9**, 1323–1332 (2022).
16. Youngblood, N., Ríos Ocampo, C. A., Pernice, W. H. P. & Bhaskaran, H. Integrated optical memristors. *Nat. Photonics* **17**, 561–572 (2023).
17. Tossoun, B. et al. High-speed and energy-efficient non-volatile silicon photonic memory based on heterogeneously integrated memresonator. *Nat. Commun.* **15**, 551 (2024).
18. Martin-Monier, L. et al. Endurance of chalcogenide optical phase change materials: a review. *Optical Mater. Express* **12**, 2145–2167 (2022).
19. Liu, Z. et al. Multilevel halide perovskite memristors based on optical & electrical resistive switching effects. *Mater. Chem. Phys.* **288**, 126393 (2022).
20. Howard, E. R., Marquez, B. A. & Shastri, B. J. Photonic long-short term memory neural networks with analog memory. In: *Proc. IEEE Photonics Conference (IPC)* 1–2 (IEEE, 2020).
21. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
22. Ma, B., Huang, Q. & Yu, F. A 12-bit 1.74-mW 20-MS/s DAC with resistor-string and current-steering hybrid architecture. In: *Proc. 28th IEEE International System-on-Chip Conference (SOCC)* 1–6 (IEEE, 2015).
23. Ansari, E. & Wentzloff, D. D. A 5mW 250MS/s 12-bit synthesized digital to analog converter. In: *Proc. IEEE 2014 Custom Integrated Circuits Conference* 1–4 (IEEE, 2014).
24. Kumar, S., Chauhan, R. K., Kumar, M. & Gupta, M. D. Impact of different technology node on the delay and power dissipation of 6T SRAM cell. In: *Proc. 7th International Conference on Signal Processing and Communication (ICSC)* 337–342 (2021).
25. Wu, X., Palmers, P. & Steyaert, M. S. J. A 130 nm CMOS 6-bit full Nyquist 3 GS/s DAC. *IEEE J. Solid-State Circuits* **43**, 2396–2403 (2008).
26. Hong, C.-H. et al. A 28nm 36kb high speed 6T SRAM with source follower PMOS read and bit-line under-drive. In: *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)* 2549–2552 (2015).
27. Widmann, D. et al. A time-interleaved digital-to-analog converter up to 118 GS/s with integrated analog multiplexer in 28-nm FD-SOI CMOS technology. *IEEE J. Solid-State Circuits* 1–15 (2023).
28. Zhang, H. et al. A 0.6-V 10-bit 200-kS/s SAR ADC with higher side-reset-and-set switching scheme and hybrid CAP-MOS DAC. *IEEE Trans. Circuits Syst. I: Regul. Pap.* **65**, 3639–3650 (2018).
29. Kumar, M. & Ubhi, J. S. Performance evaluation of 6T, 7T & 8T SRAM at 180 nm technology. In: *Proc. 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* 1–6 (IEEE, 2017).
30. Rubino, R., Musolino, F., Chen, Y., Richelli, A. & Crovetto, P. A 880 nW, 100 kS/s, 13 bit differential relaxation-DAC in 180 nm. In: *Proc. 18th Conference on Ph.D Research in Microelectronics and Electronics (PRIME)* 269–272 (IEEE, 2023).
31. den Boer, W. & Den Boer, W. *Active Matrix Liquid Crystal Displays: Fundamentals and Applications* (Elsevier Science & Technology, 2005).
32. Li, G. et al. 25Gb/s 1V-driving CMOS ring modulator with integrated thermal tuning. *Opt. Express* **19**, 20435–20443 (2011).
33. Li, H. et al. A 3-D-integrated silicon photonic microring-based 112-Gb/s PAM-4 transmitter with nonlinear equalization and thermal control. *IEEE J. Solid-State Circuits* **56**, 19–29 (2021).
34. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).
35. Tsai, H., Ambrogio, S., Narayanan, P., Shelby, R. M. & Burr, G. W. Recent progress in analog memory-based accelerators for deep learning. *J. Phys. D: Appl. Phys.* **51**, 283001 (2018).
36. Horowitz, M. 1.1 Computing’s energy problem (and what we can do about it). In: *Proc. IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* 10–14 (IEEE, 2014).
37. Fick, D. Mythic @ Hot Chips 2018 (2018).
38. Wan, W. et al. A compute-in-memory chip based on resistive random-access memory. *Nature* **608**, 504–512 (2022).
39. Ambrogio, S. et al. An analog-AI chip for energy-efficient speech recognition and transcription. *Nature* **620**, 768–775 (2023).
40. Atef, M. & Zimmermann, H. Optical receiver using noise cancelling with an integrated photodiode in 40 nm CMOS technology. *IEEE Trans. Circuits Syst. I: Regul. Pap.* **60**, 1929–1936 (2013).
41. Babar, M. B., Roberts, G. & Maram, R. A 83-GHz and 68-dB $\Omega$  TIA with 2.3 pA/ $\sqrt{\text{Hz}}$ : towards high speed and low noise optical receivers. In: *Proc. 20th IEEE Interregional NEWCAS Conference (NEWCAS)* 361–365 (IEEE, 2022).
42. Awny, A. et al. A 40 Gb/s monolithically integrated linear photonic receiver in a 0.25 $\mu\text{m}$  BiCMOS SiGe:C technology. *IEEE Microw. Wirel. Compon. Lett.* **25**, 469–471 (2015).
43. Kudzus, S., Shahani, A., Pavan, S., Shafer, D. & Tarsia, M. A 46-GHz distributed transimpedance amplifier using SiGe bipolar technology. In: *Proc. IEEE MTT-S International Microwave Symposium Digest*, 2003 2, 1387–1390 (2003).
44. Awny, A. et al. 23.5 A dual 64Gbaud 10k $\Omega$  5% THD linear differential transimpedance amplifier with automatic gain control in 0.13 $\mu\text{m}$  BiCMOS technology for optical fiber coherent receivers. In: *Proc. IEEE International Solid-State Circuits Conference (ISSCC)* 406–407 (IEEE, 2016).
45. Yu, H. et al. Using carrier-depletion silicon modulators for optical power monitoring. *Opt. Lett.* **37**, 4681–4683 (2012).
46. Tait, A. N. et al. Feedback control for microring weight banks. *Opt. Express* **26**, 26422–26443 (2018).
47. Ma, X. et al. High-density integrated photonic tensor processing unit with a matrix multiply compiler. In: *Proc. IEEE Photonics Conference (IPC)* (IEEE, 2022).
48. Tait, A. N. *Silicon Photonic Neural Networks*. Ph.D. thesis (Princeton University, 2018).
49. Al-Qadasi, M. A., Chrostowski, L., Shastri, B. J. & Shekhar, S. Scaling up silicon photonic-based accelerators: challenges and opportunities. *APL Photonics* **7**, 020902 (2022).
50. Singhal, R. et al. Enhanced regularization for on-chip training using analog and temporary memory weights. *Neural Netw.* **165**, 1050–1057 (2023).
51. Weilenmann, C. et al. Single neuromorphic memristor closely emulates multiple synaptic mechanisms for energy efficient neural networks. *Nat. Commun.* **15**, 6898 (2024).
52. Miller, K. J., Haglund, R. F. & Weiss, S. M. Optical phase change materials in integrated silicon photonic devices: review. *Optical Mater. Express* **8**, 2415–2429 (2018).
53. Buckley, S. M., Tait, A. N., McCaughan, A. N. & Shastri, B. J. Photonic online learning: a perspective. *Nanophotonics* **12**, 833–845 (2023).
54. Hughes, T. W., Minkov, M., Shi, Y. & Fan, S. Training of photonic neural networks through in situ backpropagation and gradient measurement. *Optica* **5**, 864–871 (2018).
55. DeRose, C. T. et al. Thermal crosstalk limits for silicon photonic DWDM interconnects. In: *Proc. Optical Interconnects Conference* 125–126 (IEEE, 2014).
56. Moazeni, S. et al. A 40-Gb/s PAM-4 transmitter based on a ring-resonator optical DAC in 45-nm SOI CMOS. *IEEE J. Solid-State Circuits* **52**, 3503–3516 (2017).

57. Ren, Y. et al. A continuously tunable silicon double-microring filter with precise temperature tracking. *IEEE Photonics J.* **10**, 1–10 (2018).
  58. Zhang, Y., Li, Y., Feng, S. & Poon, A. W. Towards adaptively tuned silicon microring resonators for optical networks-on-chip applications. *IEEE J. Sel. Top. Quantum Electron.* **20**, 136–149 (2014).
  59. Liu, J. et al. Silicon mode-loop Mach-Zehnder modulator with L-shaped PN junction for 0.37 V cm  $V_{\pi L}$  high-efficiency modulation. *Photonics Res.* **10**, 214–221 (2022).
  60. Yong, Z. et al. U-shaped PN junctions for efficient silicon Mach-Zehnder and microring modulators in the O-band. *Opt. Express* **25**, 8425–8439 (2017).
  61. Itokawa, H., Ohno, H., Iinuma, T. & Suguro, K. Reduction in pn junction leakage for Ni-silicided small Si islands by using improved convection annealing. *Jpn. J. Appl. Phys.* **48**, 076503 (2009).
  62. Horiuchi, M. & Ohoyu, K. Reduction of PN junction leakage current by using poly-Si interlayered SOI wafers. *IEEE Trans. Electron Devices* **42**, 876–882 (1995).
  63. Hsu, W.-C., Zhou, B. & Wang, A. X. MOS capacitor-driven silicon modulators: a mini review and comparative analysis of modulation efficiency and optical loss. *IEEE J. Sel. Top. Quantum Electron.* **28**, 1–11 (2022).
  64. Chmielak, B. et al. Pockels effect based fully integrated, strained silicon electro-optic modulator. *Opt. Express* **19**, 17212–17219 (2011).
  65. Azadeh, S. S., Merget, F., Nezhad, M. P. & Witzens, J. On the measurement of the Pockels effect in strained silicon. *Opt. Lett.* **40**, 1877–1880 (2015).
  66. Eltes, F. et al. A BaTiO<sub>3</sub>-based electro-optic Pockels modulator monolithically integrated on an advanced silicon photonics platform. *J. Lightwave Technol.* **37**, 1456–1462 (2019).
  67. Shekhar, S. et al. Roadmapping the next generation of silicon photonics. *Nat. Commun.* **15**, 751 (2024).
  68. Hui, R. & O'sullivan, M. Characterization of optical devices. In Hui, R. & O'sullivan, M. (eds.) *Fiber-Optic Measurement Techniques (Second Edition)*, 297–446 (Academic Press, 2023).
  69. Di Patrizio Stanchieri, G. et al. A 1.8 V low-power low-noise high tunable gain TIA for CMOS integrated optoelectronic biomedical applications. *Electronics* **11**, 1271 (2022).
  70. Chuah, J. H. & Holburn, D. Design of low-noise high-gain CMOS transimpedance amplifier for intelligent sensing of secondary electrons. *IEEE Sens. J.* **15**, 5997–6004 (2015).
  71. Wang, P. et al. A multi-channel low-noise analog front end circuit for linear LADAR. *IEEE Trans. Circuits Syst. II: Express Briefs* **67**, 1209–1213 (2020).
  72. Li, H., Balamurugan, G., Jaussi, J. & Casper, B. A 112 Gb/s PAM4 linear TIA with 0.96 pJ/bit energy efficiency in 28 nm CMOS. In: *Proc. IEEE Conference Publication|IEEE Xplore. ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC)* (IEEE, 2018).
  73. Geuzebroek, D. H. & Driessen, A. Ring-resonator-based wavelength filters. In Venghaus, H. (ed.) *Wavelength Filters in Fibre Optics, Springer Series in Optical Sciences*, 341–379 (Springer, 2006).
- tools. We also acknowledge Omid Esmaeeli and Mohammed Al-Qadasi for helping with design reviews, design rule checks, and tapeout; Mustafa Hammood for helping with wirebonding; Jonathan Barnes and Madeline Mahanloo for helping with measurements; Avilash Mukherjee, Hassan Taleaian, Ata Khorami, Mohammed Al-Qadasi, and Ben Cohen for reviewing this paper. S.L., S.S., L.C., and B.J.S. acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC). B.J.S. is supported by the Canada Research Chairs Program and the Sloan Foundation. S.S. is supported by the Schmidt Science Polymath Award. S.B. acknowledges funding from the Fonds de recherche du Québec-Nature et technologies.

## Author contributions

S.L., S.S., B.A.M., and B.J.S. conceived the idea. S.L. performed the experiments, led the manuscript writing. S.L., S.S., and S.B. architected, and S.L. and S.B. designed the chip. A.K. and S.L. developed and analyzed the simulation studies. S.L., A.K., S.B., B.A.M., P.R.P., L.C., B.J.S., and S.S. analyzed and discussed the data and contributed to revising the manuscript. S.S. and B.J.S. co-supervised this study.

## Competing interests

B.J.S. and B.A.M. cofounded Milkshake Technology Inc. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-69084-x>.

**Correspondence** and requests for materials should be addressed to Sean Lam, Bhavin J. Shastri or Sudip Shekhar.

**Peer review information** *Nature Communications* thanks Alessandro Catania, Bowei Dong, and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

## Acknowledgements

We gratefully acknowledge Vikas Vijay Kumar and the GlobalFoundries University Program team for their generous MPW support, and we acknowledge the Natural Sciences and Engineering Research Council of Canada and CMC Microsystems for their funding support and design