COMPUTER NETWORKS

Delocalized photonic deep learning on the internet's edge

Alexander Sludds[1]*, Saumil Bandyopadhyay[1], Zaijun Chen[1]†, Zhizhen Zhong[2], Jared Cochrane[1,3], Liane Bernstein[1], Darius Bunandar[1]‡, P. Ben Dixon[3], Scott A. Hamilton[3], Matthew Streshinsky[4]§, Ari Novack[4]§, Tom Baehr-Jones[4]§, Michael Hochberg[4]§, Manya Ghobadi[2], Ryan Hamerly[1,5]*, Dirk Englund[1]*

Advanced machine learning models are currently impossible to run on edge devices such as smart sensors and unmanned aerial vehicles owing to constraints on power, processing, and memory. We introduce an approach to machine learning inference based on delocalized analog processing across networks. In this approach, named Netcast, cloud-based "smart transceivers" stream weight data to edge devices, enabling ultraefficient photonic inference. We demonstrate image recognition at ultralow optical energy of 40 attojoules per multiply (<1 photon per multiply) at 98.8% (93%) classification accuracy. We reproduce this performance in a Boston-area field trial over 86 kilometers of deployed optical fiber, wavelength multiplexed over 3 terahertz of optical bandwidth. Netcast allows milliwatt-class edge devices with minimal memory and processing to compute at teraFLOPS rates reserved for high-power (>100 watts) cloud computers.

Advances in deep neural networks (DNNs) are transforming science and technology (*1*–*4*). However, the increasing computational demands of the most powerful DNNs limit deployment on low-power devices, such as smartphones and sensors—and this trend is accelerated by the simultaneous move toward Internet of Things (IoT) devices. Numerous efforts are underway to lower power consumption, but a fundamental bottleneck remains because of energy consumption in matrix algebra (*5*), even for analog approaches including neuromorphic (*6*), analog memory (*7*), and photonic meshes (*8*). In all these approaches, memory access and multiply-accumulate (MAC) functions remain a stubborn bottleneck near 1 pJ per MAC (*5*, *9*–*12*). Edge devices typically use chip-scale sensors, occupy millimeter-scale footprints, and consume milliwatts of power. Their small footprint and low power budget mean that performance is limited by the size, weight, and power (SWaP) of computing systems integrated on the device.

To make advanced DNNs at all feasible on low-power devices, industry has resorted to offloading computationally heavy DNN inference to cloud servers. For instance, a smart home device may send a voice query as a vector U to a cloud server, which returns the inference result V to the client (Fig. 1). This offloading architecture adds a ~200-ms latency to voice commands (*13*), which makes services such as self-driving impossible. Moreover, offloading poses security risks in both the edge and the cloud: Hacking of the communication

[1]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02421, USA. [4]Nokia Corporation, New York, NY 10016, USA. [5]Physics and Informatics Laboratories, NTT Research Inc., Sunnyvale, CA 94085, USA.
*Corresponding author. Email: asludds@mit.edu (A.S.); rhamerly@mit.edu (R.H.); englund@mit.edu (D.E.)
†Present address: Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA. ‡Present address: Lightmatter Inc., Boston, MA 02110, USA. §Present address: Luminous Computing Inc., Mountain View, CA 94041, USA.

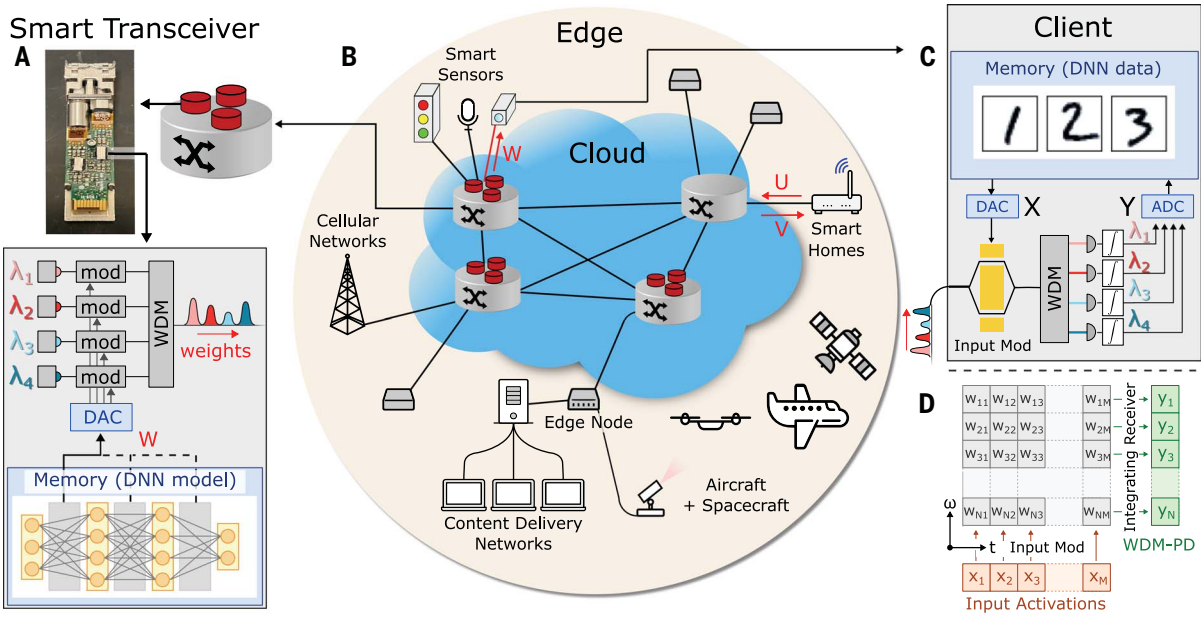

Fig. 1. Netcast concept. (**A**) Smart transceivers integrated alongside cloud computing infrastructure including servers, data storage, network switches, and edge nodes. The smart transceiver sequentially encodes layers of a neural network model onto the intensity of distinct optical wavelengths using digital-to-analog converters (DACs), optical modulators (mod), and lasers. Wavelength-division multiplexers (WDMs) combine the separate wavelengths from each modulator to the smart transceiver output. (**B**) *U* and *V* highlight current solutions to large model deployments on the edge, with edge device data communicated back to cloud computers. In our solution, smart transceivers have connections to many devices at the edge of the communications network, including cellular networks, smart sensors, content delivery networks, and aircraft. (**C**) The edge client encodes input activation data onto a single broadband optical modulator, modulating all weight wavelengths simultaneously. Wavelengths are separated with a WDM, and the result of matrix-vector multiplication is computed on time-integrating receivers. (**D**) Matrix-vector products between an *M*-element input vector and (*M*,*N*) weight matrix are time-frequency (t-ω) encoded, with each wavelength accumulating its results on a time-integrating receiver.

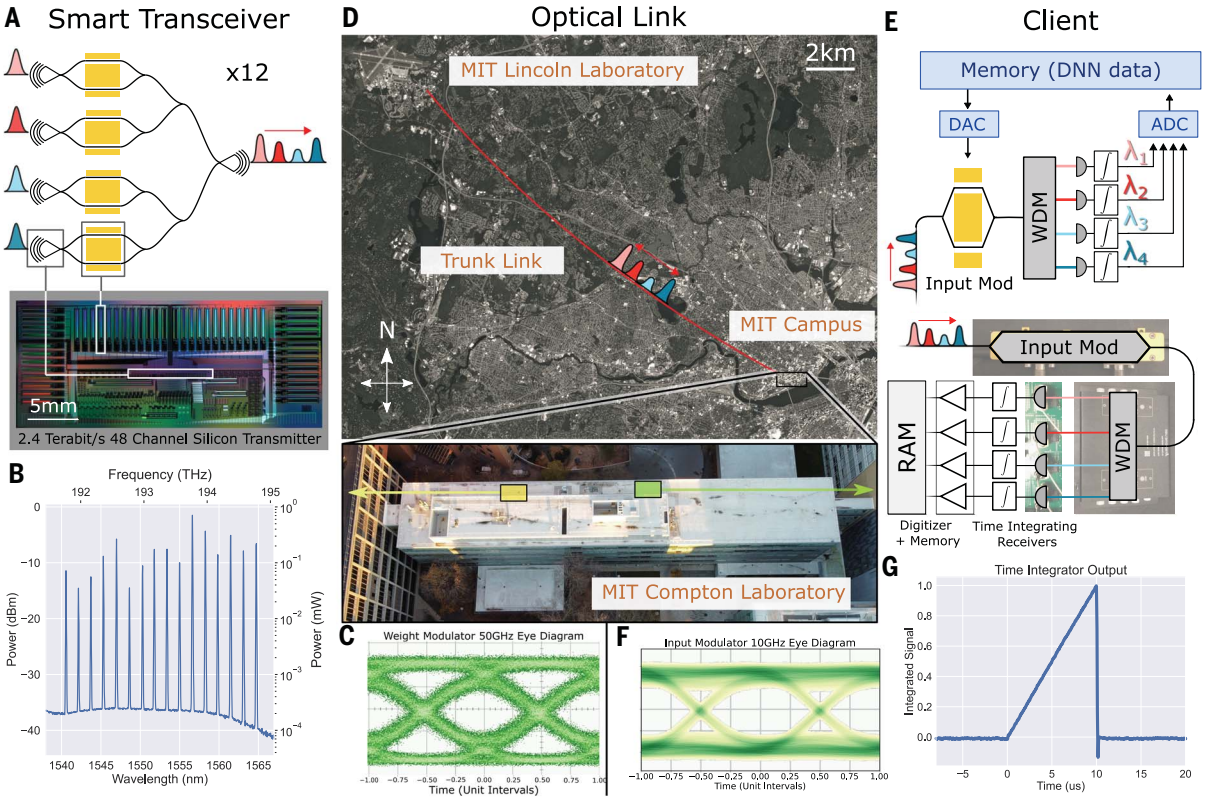

Fig. 2. Experimental demonstration of Netcast system. (**A**) Smart transceiver composed of a 48-modulator silicon photonic transmitter with 2.4 Tbps of total bandwidth. (**B**) Optical spectrum of smart transceiver output, showing 16 laser sources across 3 THz of bandwidth with >25 dB optical SNR. (**C**) An example of high-speed operation of the smart transceiver modulators, with a 50 GHz open eye. (**D**) Weights are sent over 86 km of deployed optical fiber connecting the smart transceiver to the client. (**E**) Client receiver composed of a broadband, high-speed optical modulator, a WDM demultiplexer, and custom time-integrating receivers. (**F**) The client input modulator also achieves an open eye of 10 GHz (test equipment limited). (**G**) Example time-integrating receiver waveform showing constant optical power being accumulated over 10 μs and resetting. Satellite imagery in (D) taken using a deployed satellite (Planet.com).

of client data (in vector U) has led to security breaches of private data.

To address these problems, we introduce here a photonic edge computing architecture, named Netcast, to minimize the energy and latency of large linear algebra operations such as general matrix-vector multiplication (GEMV) (*5*). In the Netcast architecture, cloud servers stream DNN weight data (W) to edge devices in an analog format for ultraefficient optical GEMV that eliminates all local weight memory access (*14*).

Servers containing a "smart transceiver" (*15*)—which may be in the standard pluggable transceiver format represented in Fig. 1A—periodically broadcast the weights (W) of commonly used DNNs to edge devices, using wavelength division multiplexing (WDM) to leverage the large spectrum available at the local access layer. Specifically, the (M,N)-sized weight matrix of one DNN layer may be encoded in a time-frequency basis by the amplitude-modulated field $W_n(t) = \sum_{j=1}^{M} w_{nj} e^{-i\omega_n t}\delta(t - j\Delta T)$, where the optical amplitude w_{nj} at frequency ω_n and time step j represents the nth row of the weight matrix (Fig. 1D), and δ is the impulse response function.

Suppose now that a camera in Fig. 1 requires inference on an image X. To do so, it waits for the server to stream the "image recognition" DNN weights, which it modulates with $X(t) = \sum_{j=1}^{M} x_j \delta(t - j\Delta T)$ using a broadband optical modulator and subsequently separates the wavelengths to N time-integrating detectors to produce the vector-vector dot product $Y_n(t) = \sum_{j=1}^{M} w_{nj} x_j \delta(t - j\Delta t)$. This architecture minimizes the active components at the client, requiring only a single optical modulator, digital-to-analog converter (DAC), and analog-to-digital converter (ADC).

Experimental implementation of Netcast

We demonstrate the Netcast protocol with a smart transceiver (Fig. 2A), made in a commercial silicon-photonic CMOS foundry (OpSIS/IME, described in supplementary text section 2). The smart transceiver is composed of 48 Mach-Zehnder modulators (MZMs), each capable of modulation up to 50 Gbps for a total bandwidth of 2.4 Tbps (*16*). The smart transceiver supports WDM, with Fig. 2B showing 16 WDM lasers simultaneously transmitting through the chip with ≈–10 dBm (100 μW) power per wavelength. Figure 2C shows an open eye diagram at 50 GHz (supplementary text section 8). Weights are transmitted over 86 km of deployed optical fiber from the Massachusetts Institute of Technology (MIT) main campus to MIT Lincoln Laboratory and back to the main campus (Fig. 2D). The client (Fig. 2E) applies input activation values to the incoming weight data using a high-speed (20-GHz) broadband lithium niobate MZM, with Fig. 2F showing an open eye diagram at 10 GHz (limited by testing equipment). A passive wavelength demultiplexer separates each wavelength channel for detection onto an array of custom time-integrating receivers, with an example of time integration shown in Fig. 2G (supplementary text section 6). After integration, the generated voltages from the receivers are measured by a digitizer and stored in memory. Additional postprocessing steps, such as the nonlinear activation function, are performed using a computer. Multiple neural network layers

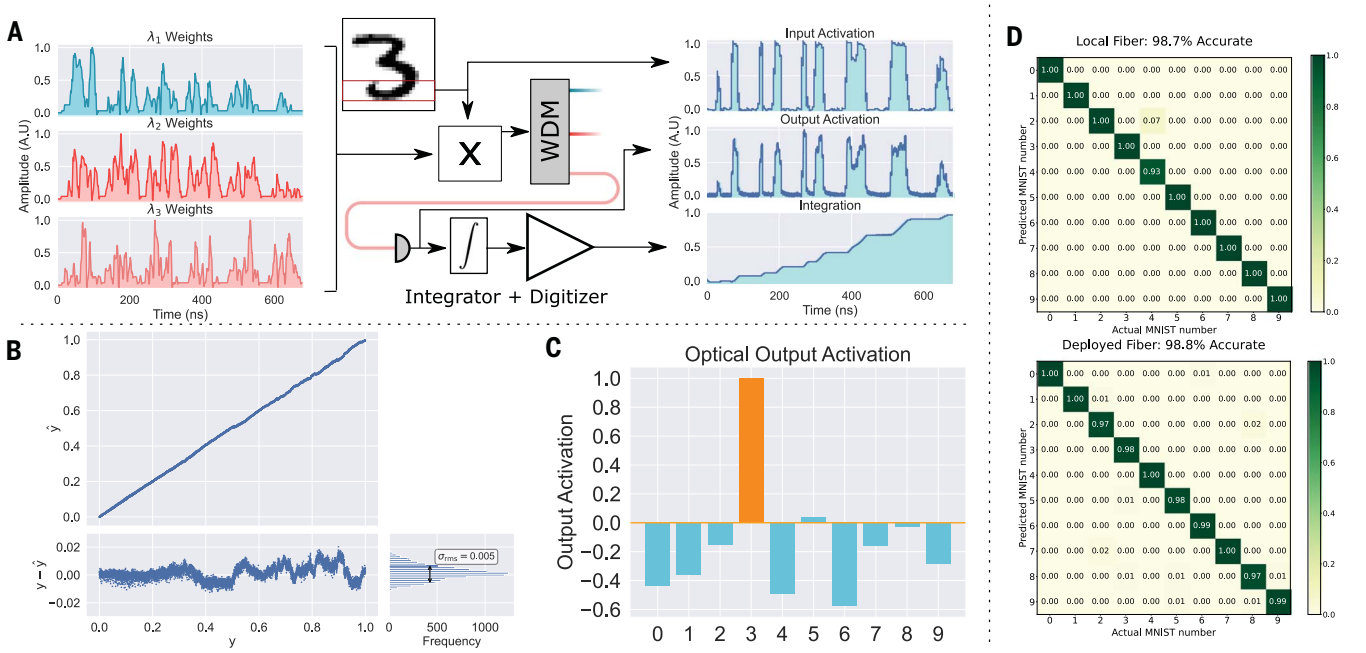

Fig. 3. Computational accuracy of Netcast system. (**A**) Weight data from multiple wavelength channels is simultaneously modulated by input data. After wavelength multiplexing, the generated photocurrent is time-integrated. (**B**) Floating-point computing accuracy comparing the results of 10,000 scalar-scalar floating point multiplications. Electrical floating point results are designated as y and optical results are designated as $\hat{y}$. The difference $y - \hat{y}$ has a standard deviation of σ_{rms} = 0.005 or ≈8-bit accuracy. (**C**) Example output activation data from the optical setup correctly classifying the digit "3." (**D**) Computing results of image classification over both local links and the 86-km deployed fiber link.

Table 1. Device contributions to receiver performance assuming conventional technology. Device energy consumption is amortized by either a spatial fan-out factor (N) or time-domain fan-out factor (M). We assume a carrier depletion modulator in silicon is used and that a single high-speed (gigahertz) ADC reads out from an array of N slow integrators. See supplementary text section 19 for derivation of nonlinearity energy consumption.

Netcast client energy consumption

Device	Number of devices	Fan-out	Energy per device	Energy per MAC
Modulator (*16*)	1	N	~1 pJ	~(1/N) pJ
DAC (*37*)	1	N	~1 pJ	~(1/N) pJ
ADC (*38*)	1	M	~1 pJ	~(1/M) pJ
Integrator (*39*)	N	M	~1 fJ	~(1/M) fJ
Nonlinearity	N	M	<100 fJ	~(1/M) fJ
Total	–	–	–	~(1/N) pJ

are run by taking the resulting output activations of the previous layer and encoding them onto the input modulator while the next layer's weights are transmitted.

We show the flow of data through the experimental setup and the accuracy it can achieve in Fig. 3A. Weight data are encoded to multiple modulators simultaneously. For clarity, we show a single row of the digit "3" being encoded and the resulting time trace from a single wavelength. We demonstrate computing with high accuracy, with Fig. 3B showing 8 bits of precision, more than the ≈5 bits of precision required for neural network computation (*17*, *18*). After calibrating the system, we perform image classification by running a benchmark handwritten digit classification task [Modified National Institute of Standards and Technology (MNIST)], which was trained on a digital computer (supplementary text sections 14 and 16). Figure 3C illustrates an example of the system's computing result for classifying the digit "3." We then test the system's performance both locally and over deployed fiber using a benchmark three-layer MNIST model with 100 neurons per hidden layer (supplementary text section 14). Using 1000 test images locally, we demonstrate 98.7% accurate computation, comparable with the model's baseline accuracy of 98.7%. Using the same test images, we utilize 3 THz of bandwidth over the deployed fiber and classify MNIST digits with 98.8% accuracy. This result shows the potential for this architecture to support ultrahigh bandwidths in real-world deployed systems using conventional components.

Energy efficiency

Netcast is designed to minimize the power used at the client. To enable this, we make sure every component at the client is performing a large number of MACs (M or N) for modulation and electrical readout, respectively. Only a single MZM and DAC are used to encode input data across N wavelengths, enabling N MACs of work for every voltage applied to the modulator. While the energy costs of these individual components can be high, they have high parallelism, performing many MACs of work per time step. For encoding input activations, the client only uses a single broadband optical modulator, allowing for ≈(1/N) pJ per MAC of energy consumption using standard components. Furthermore, the integrator and ADC can be much slower than the speed of modulated weights, because readout occurs after M timesteps. As a result, the integrator and ADC can be M times slower, decreasing the cost of electrical readout components to ≈(1/M) pJ per MAC. Assuming near-term values of N = M = 100, client energy consumption can reach

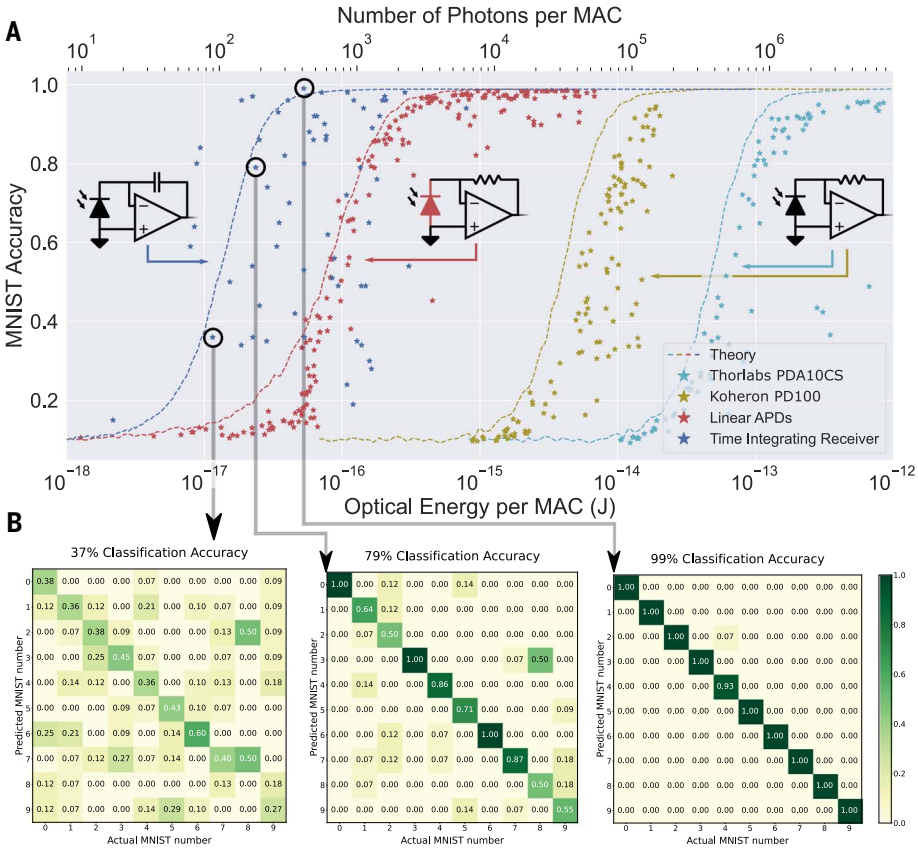

Fig. 4. Thermal noise limited optical sensitivity of Netcast system. (**A**) Experimentally measured sensitivity of optical receivers. Standard amplified photoreceivers are shown on the right side of the plot, with performance limited by electrical amplifier thermal noise, giving a typically optical energy of 10 to 100 fJ per MAC. The center of the plot shows linear avalanche photodiodes, which use intrinsic gain to lower the energy per MAC, but at the cost of increased energy consumption and lower-bandwidth time-integrating receivers, which lower the effective thermal noise floor by performing many MAC operations for each readout. Time-integrating receivers using off-the-shelf technology can achieve high accuracy with <100 aJ per MAC of optical sensitivity on the benchmark neural network task. (**B**) Confusion matrices for labeled points in (A), showing how each digit in the MNIST dataset is classified by the optical hardware (on-diagonal elements correspond to correct classification; columns add to 1, but rows do not have to).

≈10 fJ per MAC, which is three orders of magnitude lower than is possible in existing digital CMOS. The scaling of the client energy consumption is summarized in Table 1.

In our experimental demonstration, we have fabricated a 48-channel silicon smart transceiver to deploy weights to the client. The modulators used in this smart transceiver can operate at a data rate of 50 Gbps. The client uses a fiber lithium niobate modulator with a bandwidth of 20 GHz and energy efficiency of 18 pJ per bit (supplementary text section 1). Sharing this input modulator over 48 wavelengths, we find that our input modulator uses 370 fJ per MAC of energy. Simple changes to the client, such as making use of the same modulator at the client as we do at the smart transceiver (≈450 fJ per bit), would enable <10 fJ per MAC energy efficiency. Our integrating receivers have a 20 mW power consumption per channel, leading to an energy efficiency of 1 pJ per MAC and the potential to improve orders of magnitude with commercial technology (see Discussion section).

Receiver sensitivity

Applications of Netcast, including free-space deployment to drones or spacecraft, can operate in deeply photon-starved environments. For example, recent satellite optical communication demonstrations, such as NASA's Lunar Laser Communication Demonstration, have shown ≈100 Mbps communications to satellites orbiting the Moon with link losses in excess of 70 dB (*19*). To enable high-speed and energy-efficient machine learning on these deployments, optical receivers must have the lowest possible noise floor, ideally operating at the shot noise limit with ≈1 photon per MAC. Modern photoreceivers are limited by either thermal noise of readout electronics [also called Johnson-Nyquist noise (*20*)], shot noise, flicker ($1/f$) noise, or relative intensity noise of the laser; of these, for integrated optoelectronics, thermal and shot noise are dominant in Netcast (see supplementary text sections 13 and 23). We overcome this problem with time-integrating receivers, which accumulate partial results from vector-matrix multiplication. We compare the sensitivity of different photoreceivers. Amplified photoreceivers (Fig. 4A, right) have typical sensitivities of ≈10 to 100 fJ per MAC. Amplified linear mode avalanche photodetectors (Fig. 4A, middle) overcome some of the thermal noise of the amplifier and achieve ≈1 fJ per MAC. Our custom time-integrating receivers (Fig. 4A, left) perform M MACs per measurement window before readout, lowering the required optical power per readout by M. Amplified photodetectors, in contrast, read out after each MAC, acquiring thermal noise for each measurement and adding the results of each MAC together to create the resulting output activation value. For time-integrating receivers, the resulting output activation signal is measured while measuring thermal noise once, giving a $\frac{1}{M}$ optical energy per MAC scaling. For amplified photodetectors, the partial-product signal terms add together linearly, while thermal noise adds in quadrature, giving a $\frac{\sqrt{M}}{M} = \frac{1}{\sqrt{M}}$ scaling. In our experiment, we demonstrate that with M = 100, only 10 aJ per MAC (100 photons) of optical energy is required (two orders of magnitude less than for similar amplified photodetectors). This result brings Netcast close to the fundamental quantum limit of optical computation (*21*, *22*), which we can reach by engineering the receiver to lower thermal noise.

Thermal noise is a hardware-dependent noise source, originating from the thermal motion of charge carriers in an electrical conductor. In a resistor-capacitor (RC) circuit, thermal noise manifests in a fluctuation in the number of readout electrons in a circuit given by $\sigma_{th} = \sqrt{k_B TC}/q$, where k_B is the Boltzmann constant, T is temperature, q is the electron charge, and C is the capacitance of the receiver (*23*). Conventional amplified photodetectors read out on every MAC operation and add partial-product results to generate an output activation value. Adding together each MAC adds together the measured signal linearly, and noise terms add in quadrature. This results in a signal-to-noise ratio (SNR) of

$$\begin{aligned}\mathrm{SNR} &= \frac{\text{Signal Electrons}}{\text{Noise Electrons}} \\ &= MP_{opt}\eta T_{clk}/\sqrt{\sum_{i=1}^{M}\sigma_{th}^2} \\ &= MP_{opt}\eta T_{clk}/\sqrt{M\sigma_{th}^2} \\ &= \sqrt{M}P_{opt}\eta T_{clk}/\sigma_{th}\end{aligned}$$

where P_{opt} is photon flux incident on the detector (units of photons per second), η is detector quantum efficiency, and T_{clk} is the time period for each MAC. In contrast, our

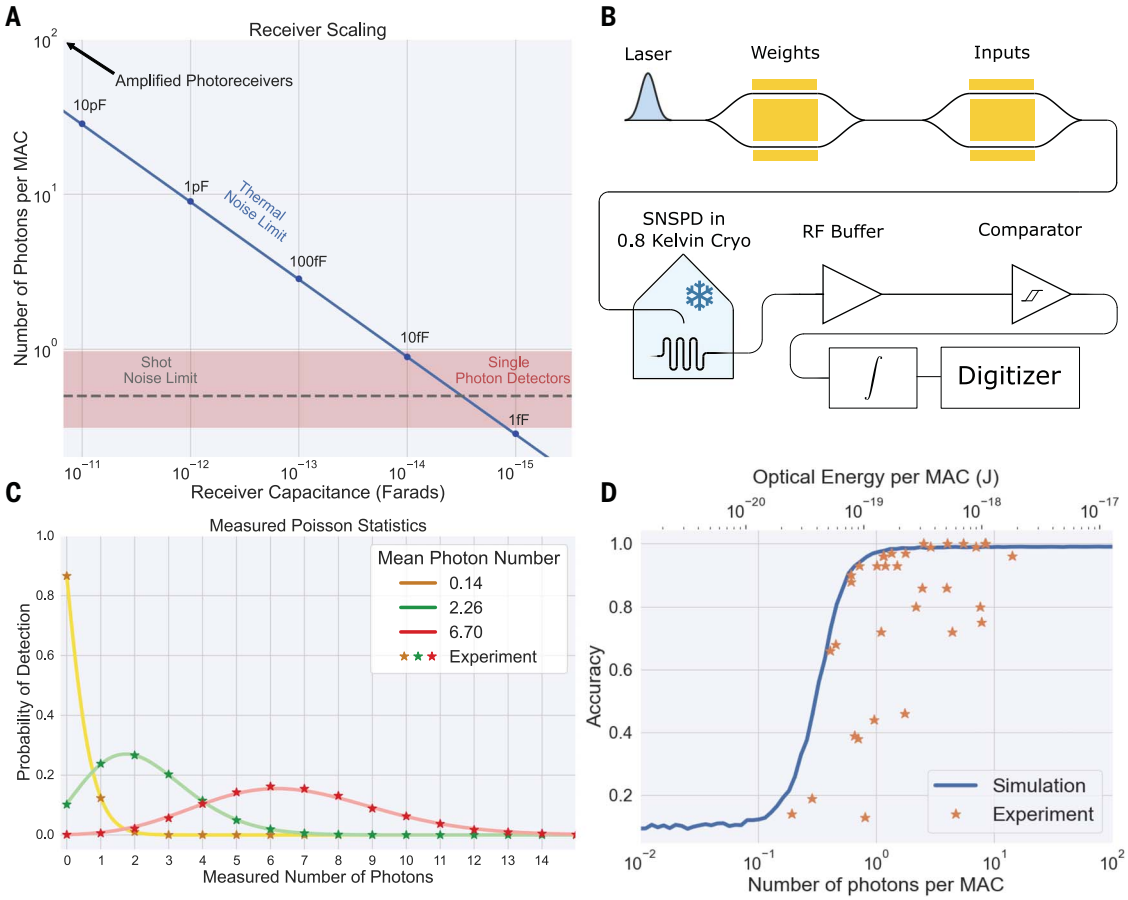

Fig. 5. Forward looking performance of Netcast. (**A**) Fundamental noise bounds of time-integrating receivers from thermal noise of an integrator and shot noise to achieve 50% accuracy on MNIST task. Decreasing the capacitance of the time integrator lowers thermal readout noise, enabling access to the single photon–per-MAC regime. (**B**) Proof-of-concept experimental setup consisting of input and weight modulators and superconducting nanowire single-photon detectors (SNSPDs), allowing us to probe this fundamental single-photon bound. (**C**) We experimentally validate the single-photon detectors by measuring shot noise on the detector over many integration windows. (**D**) Using a three-layer MNIST model, we experimentally measure computation with <1 photon per MAC with high accuracy.

time-integrating receivers only see thermal noise once per measurement window

$$\text{SNR} = \frac{\text{Signal Electrons}}{\text{Noise Electrons}} = MP_{\text{opt}}\eta T_{\text{clk}}/\sigma_{\text{th}}$$

As a result, we see that the required number of photons per MAC is $\sqrt{M}$ times lower than for standard amplified photoreceivers.

Improvements to time-integrating receivers are possible by minimizing the integration capacitance of the receiver. Figure 5A shows the thermal noise limit of time-integrating receivers as integration capacitance is decreased. This noise floor is fundamentally connected to the size scale of photodetectors, readout electronics, and their proximity of integration (*10*). Modern foundry processes enable ≈1 fF–scale receivers, lowering the thermal readout noise to the single photon–per-MAC level (*24*, *25*). This single photon–per-MAC regime is fundamentally limited by the quantum nature of light, where precision is determined by the Poissonian distribution of photons that arrive within a measurement window. Poissonian noise, also called shot noise, can be observed in experimentally measured data in Fig. 5C. We investigate this fundamental bound of the Netcast system through a proof-of-concept experiment using superconducting nanowire single-photon detectors (SNSPDs) as shown in Fig. 5B. These photodetectors are ideal, demonstrating pure shot noise–limited performance. We show that the fundamental shot noise bound on the same benchmark digit classification problem from Fig. 4 allows the receiver to operate with high accuracy with <1 photon per MAC (0.1 aJ per MAC). This result may at first seem surprising given that less than a single photon per MAC is counterintuitive. We can understand this measurement better by noting that at readout, we have performed a vector-vector product with M = 100 MACs. Each MAC can have less than a single photon in it, but the measured signal will have many photons in it. A graphical explanation is given in supplementary text section 18. This single photon–per-MAC regime enables many new applications. The realization of computing with less than one photon per MAC could enable a new class of computing systems that protect both client input and server weight data. Another application that benefits from less than one photon per MAC is deployed spacecraft that operate in a strongly photon-starved environment. Weight data from a directional base station could be transmitted to the spacecraft and classified on the craft, before the results are transmitted to Earth.

Discussion

The system-level demonstration shown here is one example of an implementation of Netcast. The cloud-based smart transceiver can reside inside of existing networking hardware such as network switches, servers, or edge nodes. Our ideas can be extended to the case where the user data are streamed through programmable network switches with smart transceivers, enabling in-network optical inference (*15*). Modern network switches, such as Intel's Tofino switch, are an ideal platform for developing Netcast commercially, as they are programmable, enabling multiple streams of weights

to be deployed at line rate (100 Gbps), and can support 64 GB of memory, reaching the storage requirements of modern neural networks. Prior work has demonstrated the feasibility of using programmable switches to perform layer-by-layer inference with smart transceivers (*15*). The large data storage of these network switches enables multiple models to be stored and queried. The client device could use its broadband modulator to allow for reflection-mode communication back to the server, where the client modulates received light and sends it back along the fiber link for communication. This querying communication can be slow and lossy, as only a few bits are required to request that a new model be sent.

Emerging photonic technologies, such as low-power static phase shifters (*26*–*28*) and high-speed phase shifters (*29*–*32*), can reduce receiver electrical energy consumption to ≈10 aJ per MAC. This energy can be further decreased by making use of the tight integration of transistors and photonics in silicon using technologies such as receiverless detectors (*10*), photonic DACs (*33*), and photonic ADCs (*34*). Detectors such as avalanche detectors could be incorporated with a time integrator to provide a benefit to the optical sensitivity of the receiver, but at the cost of added electrical power consumption (supplementary text section 21). Further improvements in optical sensitivity are possible by using coherent detection, which boosts the received signal using a strong local oscillator (*21*). Two examples of a Netcast system using coherent detection to substantially improve optical energy per MAC are detailed in supplementary text section 12.

A number of companies have designed custom edge computing application-specific integrated circuits (ASICs) with reduced SWaP (*7*, *35*), but these ASICs are hampered by the same energy and bandwidth constraints as larger CMOS processors. Analog accelerators, such as memristive crossbar arrays and meshes of photonic interferometers, hold promise for lowering the power consumption of neural networks compared with electronic counterparts, but existing commercial demonstrations still consume watts of power (*8*, *36*).

One obstacle to scaling bandwidth in traditional optical communication systems is dispersion in optical fiber. For a single smart transceiver and client, techniques such as wavelength-dependent delays can compensate for dispersion at the smart transceiver. However, in systems where weights are deployed to multiple clients from one smart transceiver with different lengths of fiber, this technique cannot be used. We discuss the effects of dispersion in supplementary text section 22 and show that it is possible to make use of the optical O-band to enable terahertz of bandwidth at clock rates of 10 GHz per wavelength over more than 10 km of optical fiber.

Outlook

We have described an edge computing architecture that makes use of the strengths of photonics and electronics to achieve orders of magnitude in energy efficiency and optical sensitivity improvements over existing digital electronics. We have demonstrated scalable photonic edge computing using WDM, time-integrating receivers, scalability to milliwatt-class power consumption, <1 photon–per-MAC receiver sensitivity, and computing over deployed fiber using 3 THz of bandwidth. On image classification tasks, we show 98.8% accurate image classification. The hardware shown in this paper is readily mass-producible from existing CMOS foundries, allowing for near-term impact on our daily lives. Our approach removes a fundamental bottleneck in edge computing, enabling high-speed computing on deployed sensors and drones.

REFERENCES AND NOTES

1. T. B. Brown *et al*., arXiv:2005.14165 [cs.CL] (2020).
2. O. Vinyals *et al*., *Nature* **575**, 350–354 (2019).
3. A. Krizhevsky, I. Sutskever, G. E. Hinton, *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
4. J. Deng *et al.*, *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 248–255.
5. V. Sze, Y.-H. Chen, T.-J. Yang, J. S. Emer, *Proc. IEEE* **105**, 2295–2329 (2017).
6. M. Davies *et al*., *IEEE Micro* **38**, 82–99 (2018).
7. Mythic, M1076 Analog Matrix Processor; https://mythic.ai/products/m1076-analog-matrix-processor/.
8. C. Demirkiran *et al*., arXiv:2109.01126 [cs.AR] (2021).
9. M. Horowitz, *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (IEEE, 2014), pp. 10–14.
10. D. A. Miller, *J. Lightwave Technol.* **35**, 346–396 (2017).
11. L. Bernstein *et al*., *Sci. Rep.* **11**, 3144 (2021).
12. Y.-H. Chen, T. Krishna, J. S. Emer, V. Sze, *IEEE J. Solid-State Circuits* **52**, 127–138 (2016).
13. M. Satyanarayanan, *Computer* **50**, 30–39 (2017).
14. R. Hamerly *et al*., *Proc. SPIE* **11804**, 118041R (2021).
15. Z. Zhong *et al.*, *Proceedings of the ACM SIGCOMM 2021 Workshop on Optical Systems (OptSys '21)* (Association for Computing Machinery, 2021), pp. 18–22.
16. M. Streshinsky *et al*., *J. Lightwave Technol.* **32**, 4370–4377 (2014).
17. T. Gokmen, M. J. Rasch, W. Haensch, *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019), pp. 22.3.1–22.3.4.
18. S. Garg, J. Lou, A. Jain, M. Nahmias, arXiv:2102.06365 [cs.LG] (2021).
19. D. M. Boroson, J. J. Scozzafava, D. V. Murphy, B. S. Robinson, M. I. T. Lincoln, *2009 Third IEEE International Conference on Space Mission Challenges for Information Technology* (IEEE, 2009), pp. 23–28.
20. J. B. Johnson, *Phys. Rev.* **32**, 97–109 (1928).
21. R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, D. Englund, *Phys. Rev. X* **9**, 021032 (2019).
22. T. Wang *et al*., *Nat. Commun.* **13**, 123 (2022).
23. J. Pierce, *Proceedings of the IRE* **44**, 601–608 (1956).
24. M. Rakowski *et al.*, *2020 Optical Fiber Communication Conference (OFC)*, OSA Technical Digest (Optica Publishing Group, 2020), paper T3H–3.
25. C. Sun *et al*., *IEEE J. Solid-State Circuits* **51**, 893–907 (2016).
26. G. Liang *et al*., *Nat. Photonics* **15**, 908–913 (2021).
27. R. Baghdadi *et al*., *Opt. Express* **29**, 19113–19119 (2021).
28. M. Dong *et al*., *Nat. Photonics* **16**, 59–65 (2022).
29. E. Timurdogan *et al*., *Nat. Commun.* **5**, 4008 (2014).
30. C. Wang *et al*., *Nature* **562**, 101–104 (2018).
31. M. Xu *et al*., *Optica* **9**, 61 (2022).
32. W. Heni *et al*., *Nat. Commun.* **10**, 1694 (2019).
33. S. Moazeni *et al*., *IEEE J. Solid-State Circuits* **52**, 3503–3516 (2017).
34. A. Zazzi *et al*., *IEEE Open J. Solid State Circuits Soc.* **1**, 209–221 (2021).
35. A. Yazdanbakhsh, K. Seshadri, B. Akin, J. Laudon, R. Narayanaswami, arXiv:2102.10423 [cs.LG] (2021).
36. D. Fick, M. Henry, "Analog computation in flash memory for datacenter-scale AI inference in a small chip," *Hot Chips 2018 (HC30)*, Cupertino, California, 19–21 August 2018.
37. B. M. Pietro Caragiulo, C. Daigle, B. Murmann, Dac performance survey 1996-2020, GitHub (2022); https://github.com/pietro-caragiulo/survey-DAC.
38. B. Murmann, ADC performance survey 1997-2021, Stanford University (2022); http://web.stanford.edu/~murmann/adcsurvey.html.
39. E. Yang, T. Lehmann, "High gain operational amplifiers in 22 nm CMOS," *2019 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2019).
40. A. Sludds, alexsludds/Delocalized_Photonic_Deep_Learning_on_the_Internets_Edge: Zenodo Added, Zenodo (2022); https://doi.org/10.5281/zenodo.6982196.

ACKNOWLEDGMENTS

We acknowledge D. Lewis and A. Pennes for assistance in machining laboratory equipment and E. Allen for discussions related to using squeezed light to further reduce photon counts. We are grateful to E. Bersin and B. Dixon for assistance in coordinating usage of the deployed fiber and A. Rizzo for help in converting and plotting eye diagram data as well as proofreading the manuscript. We thank C. Panuski and S. Krastanov for informative discussions on single-photon operation of Netcast. We thank A. Pyke of Micro-Precision Technologies for wire bonding the electrical connections from the printed circuit board to the 48-channel transmitter. We appreciate help from F. Wong for the use of his SNSPDs and acknowledge M. Prabhu and C. Errando Herranz for facilitating the usage of the SNSPDs, C. Freeman for help in taking drone photography, NVIDIA for supplying a Tesla K40 GPU that was used for simulations shown in the main text and supplementary materials, and Planet.com for allowing us to take custom satellite imagery. **Funding:** A.S. and S.B. are supported by National Science Foundation Graduate Research Fellowship 1745302. L.B. is supported by the Natural Sciences and Engineering Research Council of Canada (PGSD3-517053-2018). This research was funded by a collaboration with NTT-Research and NSF Eager (CNS-1946976). This material is based on research sponsored by the Air Force Office of Scientific Research (AFOSR) under award FA9550-20-1-0113, the Air Force Research Laboratory (AFRL) under agreement FA8750-20-2-1007, the Army Research Office (ARO) under agreement W911NF-17-1-0527, NSF RAISE-TAQS grant 1936314, NSF C-Accel grant 2040695, NSF grant ASCENT-2023468, NSF grant CAREER-2144766, and ARPA-E grant ENLITENED PINE DE-AR0000843. The work was further supported by funding from the Alfred P. Sloan foundation (FG-2022-18504) and DARPA grant FastNICs 4202290027. Distribution Statement A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering. **Author contributions:** A.S. created the experimental setup and conducted the experiment. S.B. assisted in fiber-to-chip coupling and discussions on the project. Z.C. assisted with high-speed measurements of the setup and discussions. M.S., A.N., T.B.-J., and M.H. designed and taped out the smart transceiver. D.B. packaged the 48-channel silicon transceiver. J.C. packaged the time-integrating receivers and assisted in calibration. D.E. established the fiber link between MIT and MIT Lincoln Laboratory, and S.A.H. and P.B.D. helped with its use. L.B. helped in discussion of fundamental noise sources. M.G. and Z.Z. assisted with discussions on modern telecommunication networks. R.H. conceived of the project idea. S.B., Z.C., L.B., M.S., A.N., T.B.-J., M.H., Z.Z., J.C., S.A.H., P.B.D., and M.G. provided feedback on the manuscript. A.S., D.E., and R.H. wrote the manuscript. **Competing interests:** M.H. is president of Luminous Computing. A.N. is vice president of system hardware design at Luminous Computing. T.B.-J. is vice president of engineering at Luminous Computing. M.S. is vice president of packaging, photonics, and mixed-signal at Luminous Computing. M.H., A.N., T.B.-J., and M.S. own stock in Luminous Computing. D.B. is chief scientist at Lightmatter and holds stock in the company. A.S. has interned at Lightmatter, receiving a wage. D.E. is an adviser to and holds shares in Lightmatter but received no support for this work.

S.B. has received consulting fees from Nokia Corporation. M.G. owns stock in companies with telecommunication interests (Google and Microsoft). R.H. and D.E. have filed a patent related to Netcast: PCT/US21/43593. M.G., Z.Z., L.B., A.S., R.H., and D.E. have filed a provisional patent related to Netcast: 63/191,120. The other authors declare that they have no competing interests. **Data and materials availability:** Data required to recreate results in the main text are available in Zenodo (*40*). **License information:**

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abq8271

Materials and Methods
Supplementary Text
Figs. S1 to S26
Table S1
References (*41*–*88*)

Submitted 3 May 2022; accepted 23 September 2022
10.1126/science.abq8271