A Photonic Neural Network Using < 1 Photon per Scalar Multiplication

Tianyu Wang[1], Shi-Yuan Ma[1], Logan G. Wright[1, 2], Tatsuhiro Onodera[1, 2], Brian C. Richard[3], and Peter L. McMahon[1]

[1] School of Applied and Engineering Physics, Cornell University, Ithaca, NY 14853, USA
[2] NTT Physics and Informatics Laboratories, NTT Research, Inc., Sunnyvale, CA 94085, USA
[3] School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA

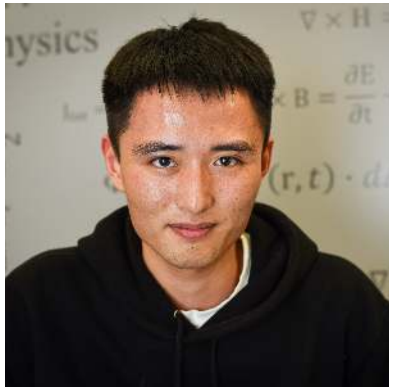
Aug 22-24, 2021

arXiv: 2104.13467

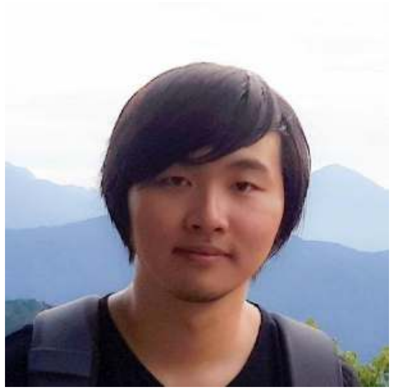
Acknowledgements

Contributors


Tianyu Wang


Shi-Yuan Ma


Logan Wright


Tatsuhiro Onodera


Brian Richard


Peter McMahon

Funding Agencies

NTT

NSF

CORNELL UNIVERSITY FOUNDED A.D. 1865

Website: https://mcmahon.aep.cornell.edu/index.html

arXiv: 2104.13467

New Frontiers in Optical Technology

The proliferation of optical sensors and user interfaces offers opportunities for deeper integration of sensing, processing, and data transmission.

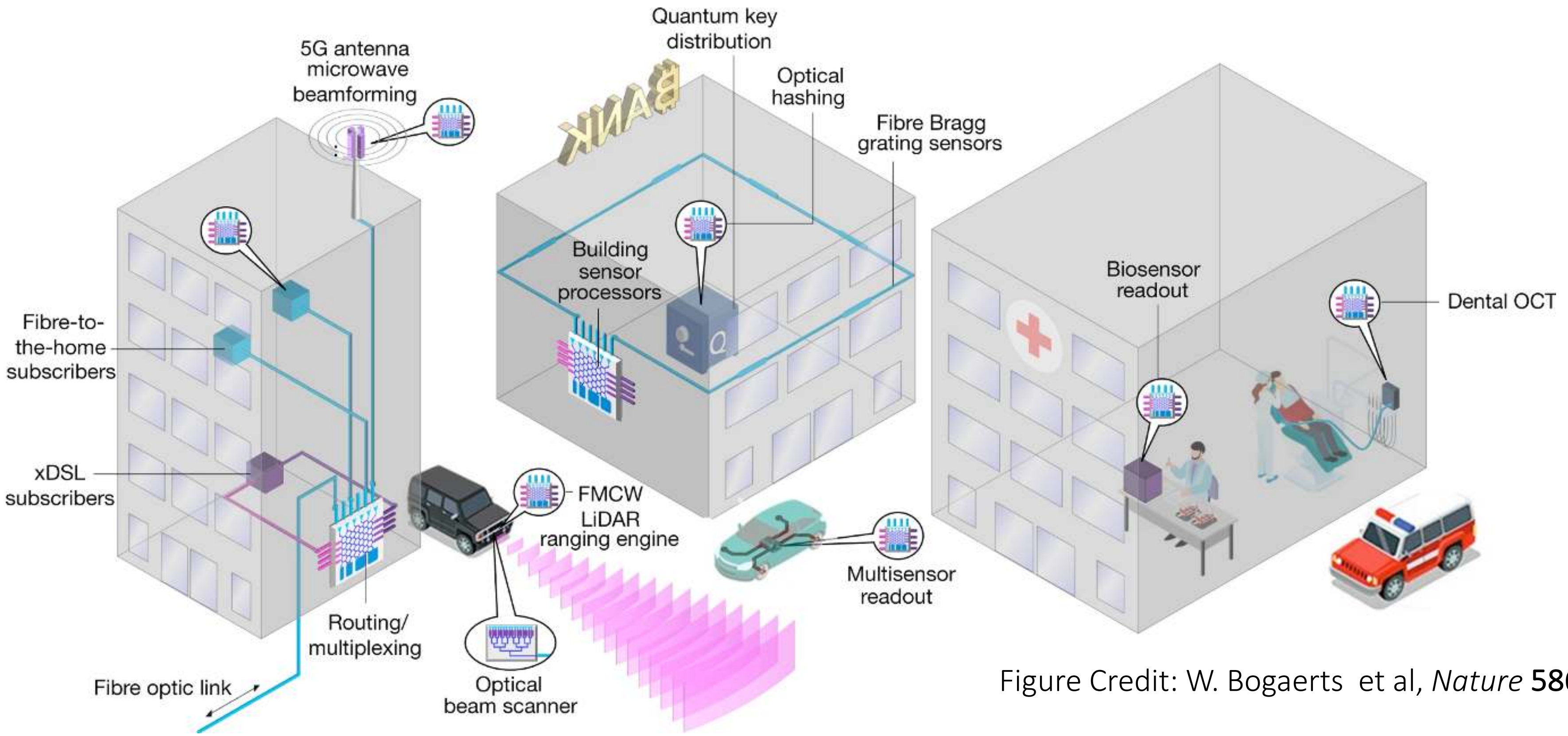


Figure Credit: W. Bogaerts et al, *Nature* **586**, 207-216 (2020).

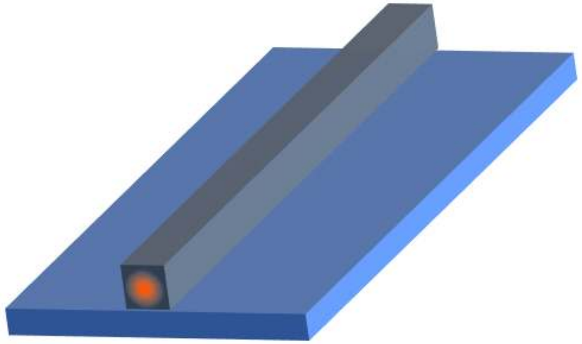
Role of optics: communication and sensing → co-processing with electronics

Advantages of Optical Processing

Optics has potentials for high-throughput parallel processing:

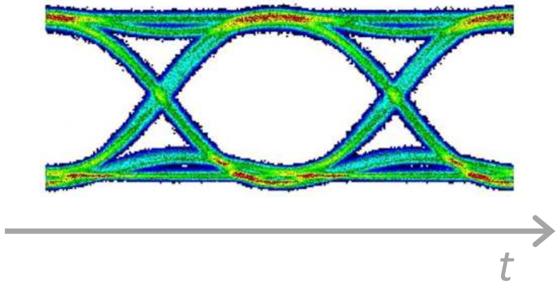
Low Transmission Loss*

0 dB in air/bulk material
~1 dB/cm in waveguides**

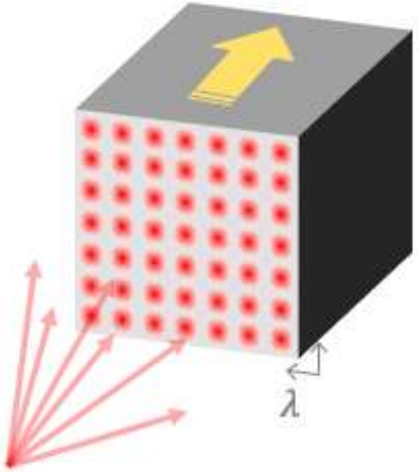


Massive Parallelism

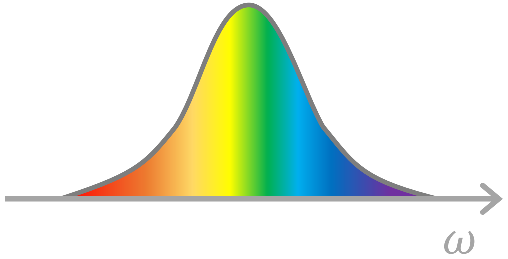
Telecom data rate
~10^{11} bit/s



Space multiplex***
~10^{6}



Wavelength multiplex
~10^{3}



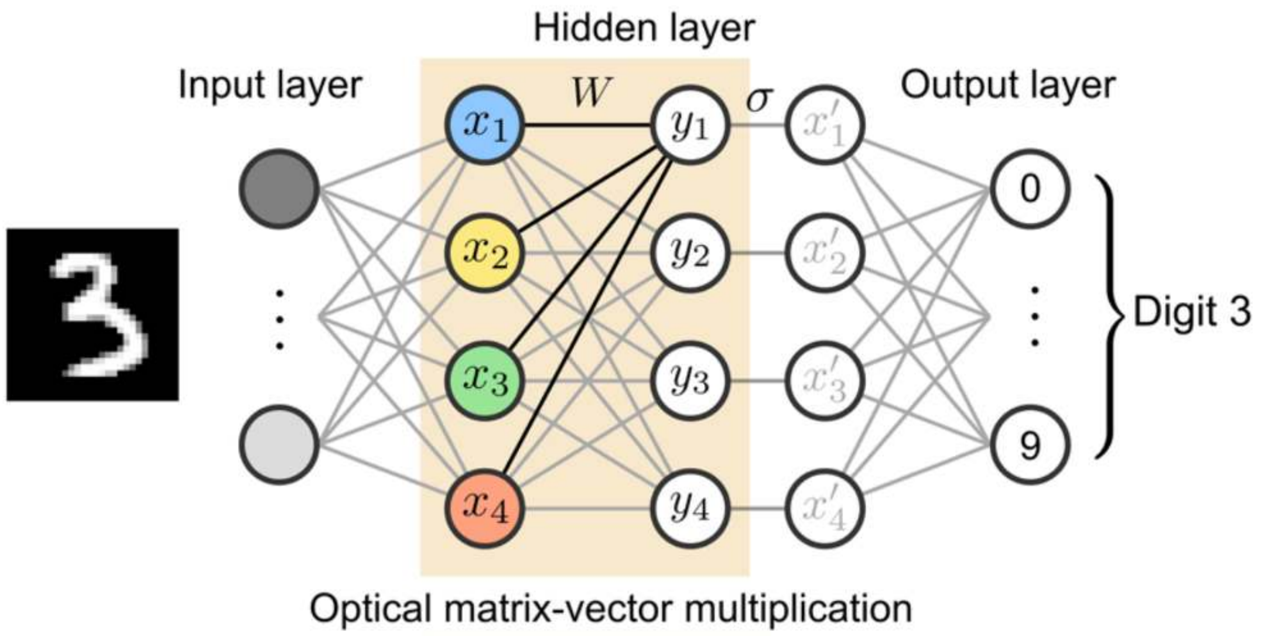
* Comparison between optical and electric transmission energy: D. A. B. Miller. *J. Light. Technol.* **35**, 346–396 (2017).

** A summary on optical waveguide performance: Su, Y., et al, Adv. Mater. Technol. **5**, 1901153, (2020).

*** Quantification of the communication capacity of optical spatial modes: D. A. B. Miller. Adv. *Opt. Photon.* **11**, 675-825 (2019).

Photonic Neural Networks (PNNs)

- Matrix-vector Multiplication (MVM) is a basic building block in deep neural networks.
- Photonic MVM can potentially achieve speed and energy benefits over electronics.



The update equation for the forward propagation in a fully connected layer:

$$x'_i = \sigma\left(\sum_{j}^{N} W_{ij} x_j + b_i\right)$$

$$O(N) \quad O(N^2) \quad O(N)$$

W: weight matrix

b : bias terms

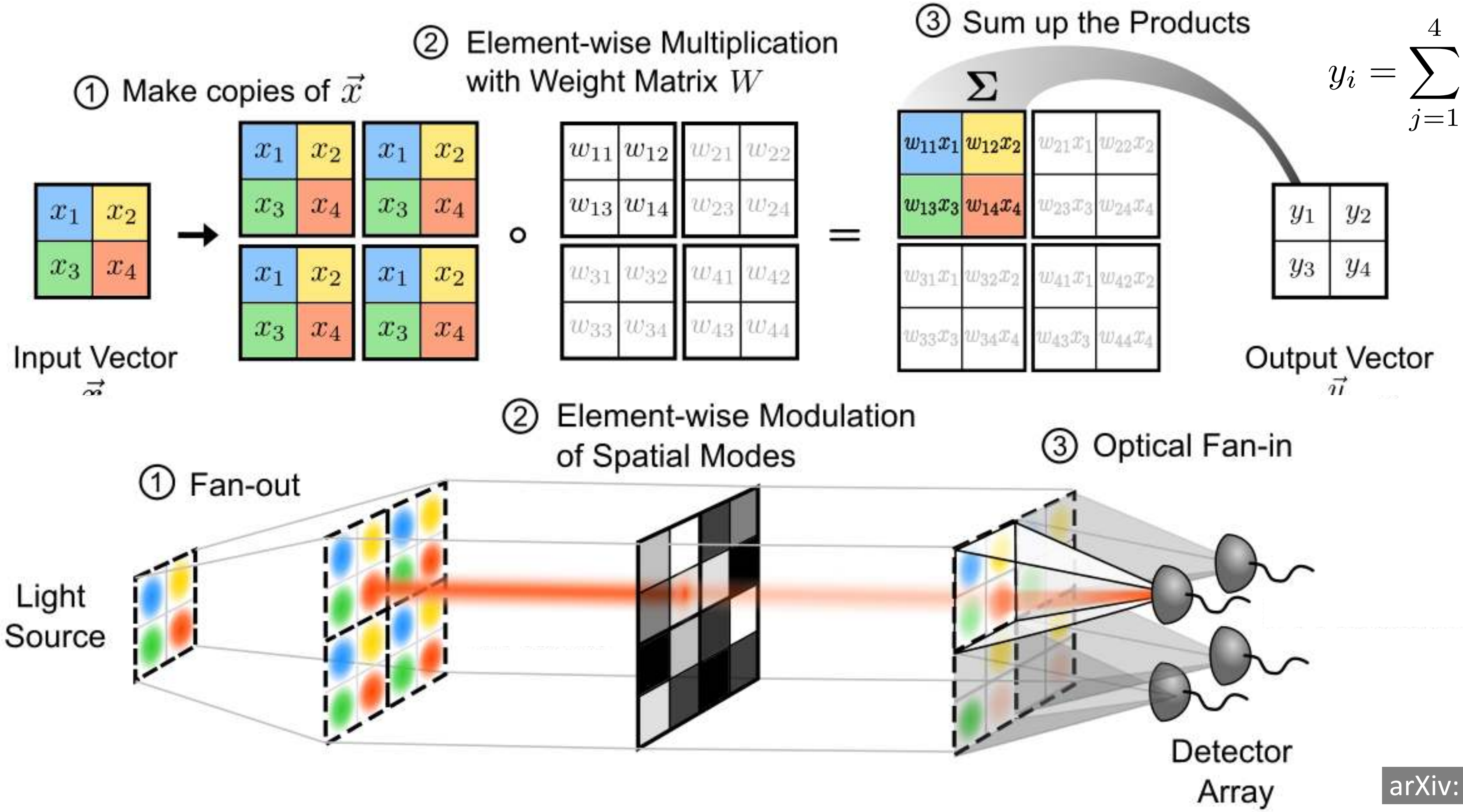
σ : nonlinear activation function

Reviews on different PNNs:
B. J. Shastri et al. *Nat. Photonics,* **15**, 102-114 (2021).
G. Wetzstein et al. *Nature*, **588**, 39-47 (2020).

arXiv: 2104.13467

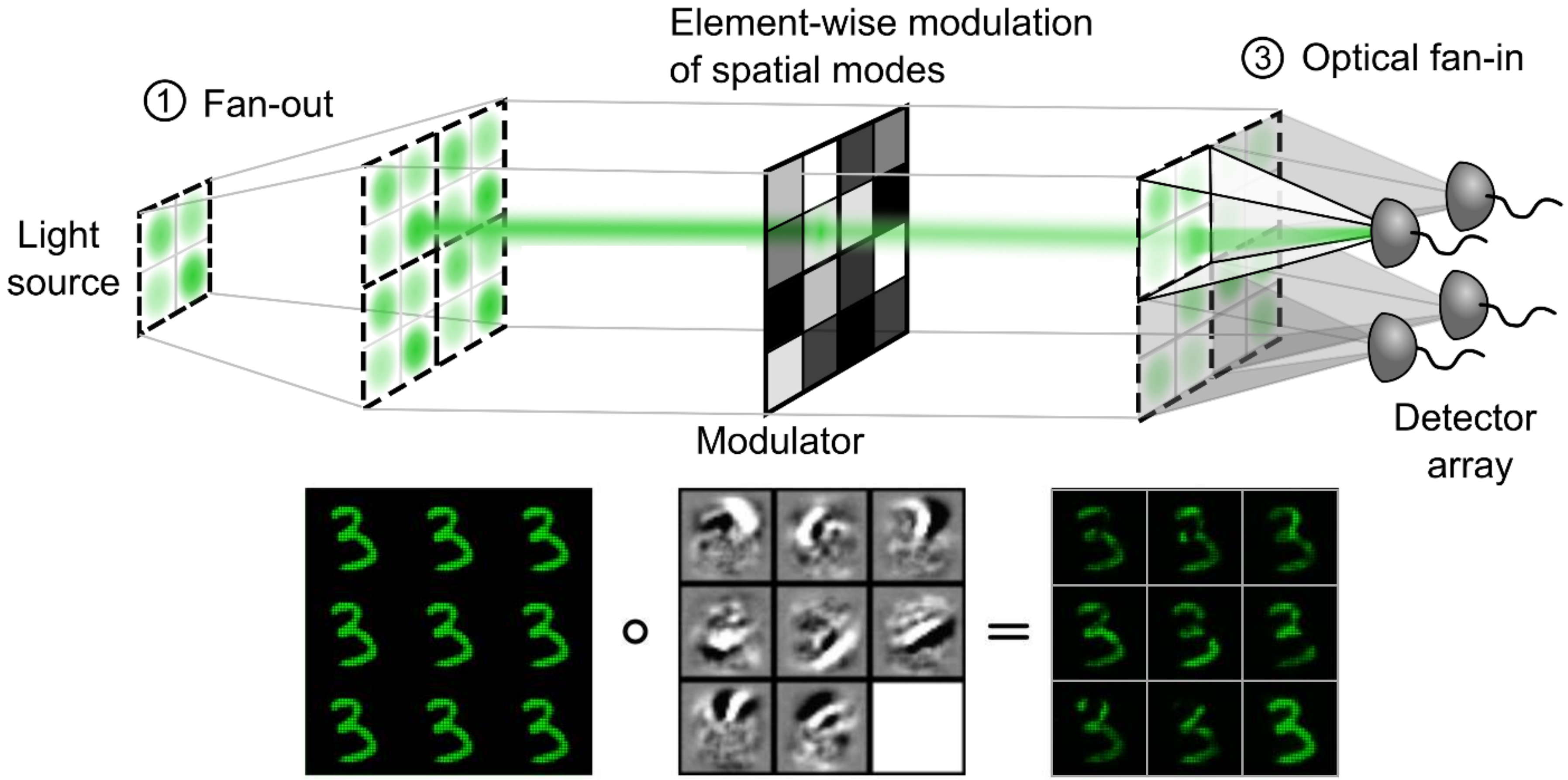
Photonic MVM Based on Free-space Optical Imaging

- To study the energy efficiency of PNNs, we constructed an optical matrix-vector multiplier that can perform generic MVM in 3 steps:



$$y_i = \sum_{j=1}^{4} w_{ij} x_j \ (i = 1, 2, 3, 4)$$

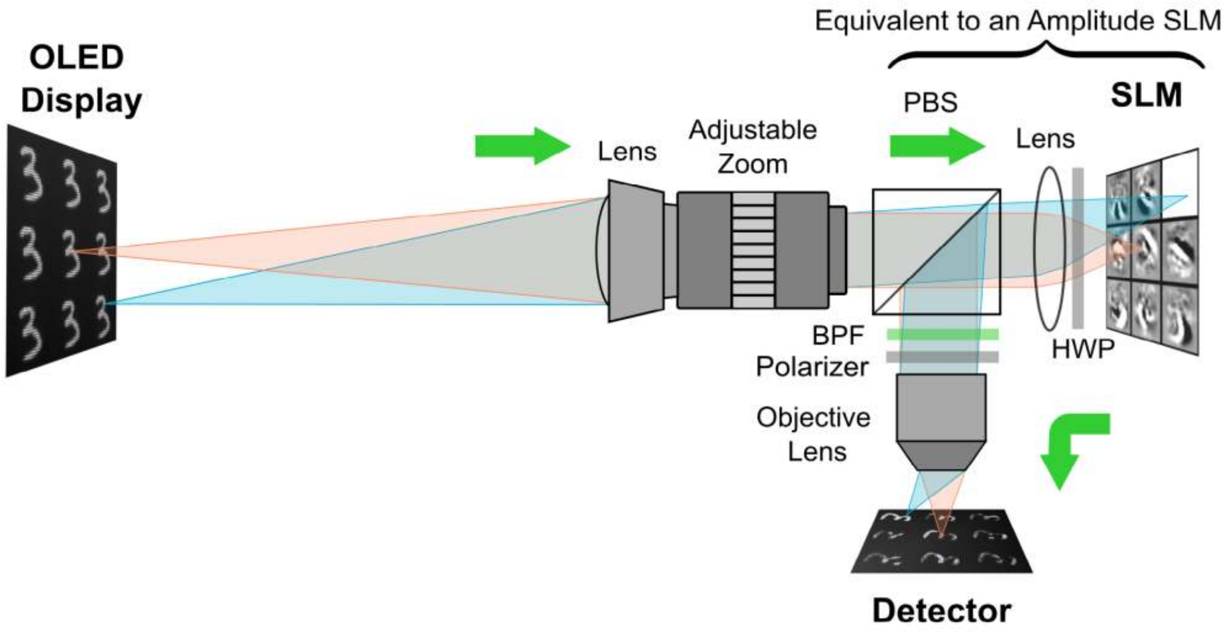
Experiment Illustration



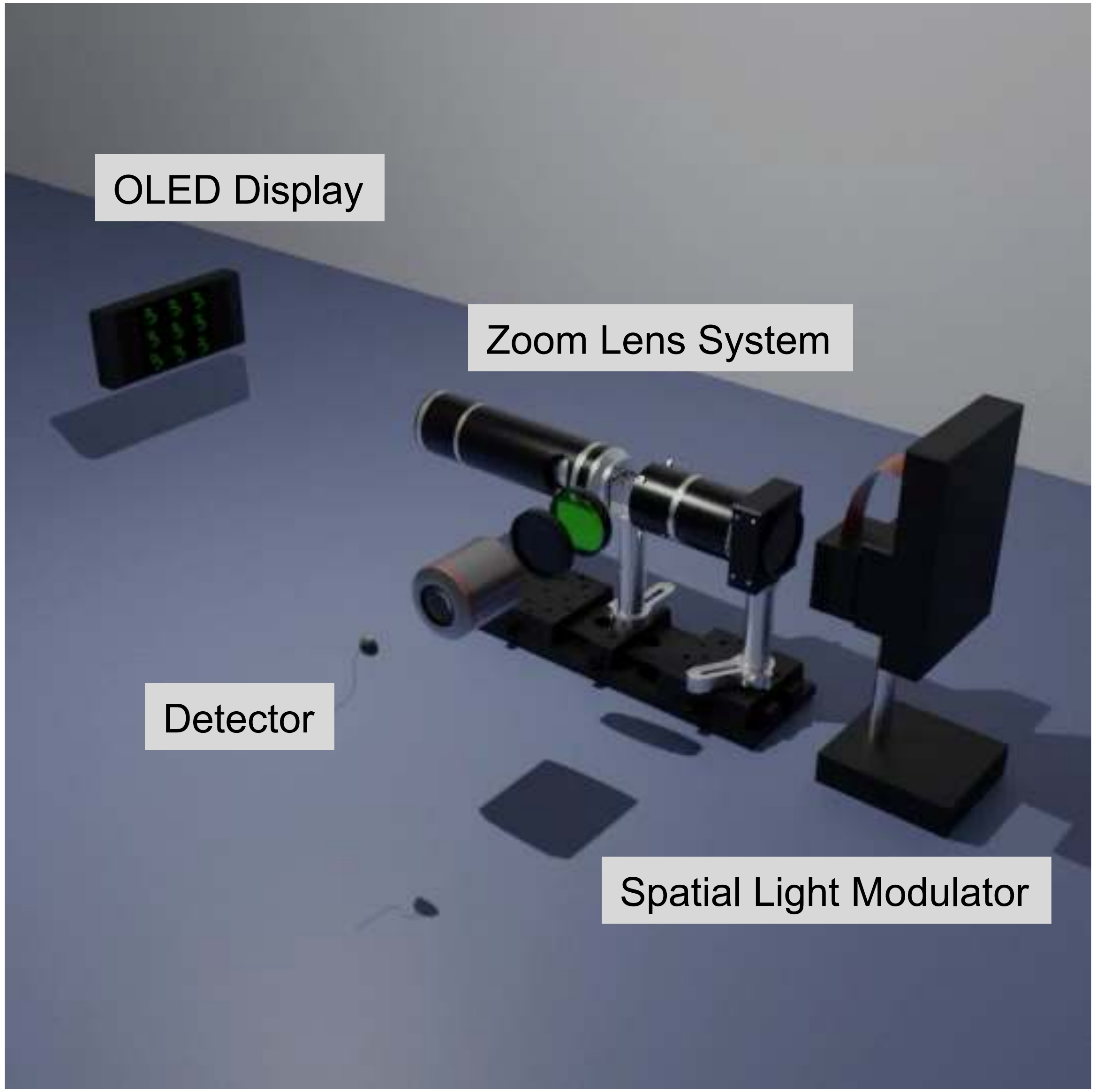
Each element of $\vec{x}$ is encoded as the intensity of a light source pixel, and W as the light transmission of a modulator pixel. Negative elements in the matrix (vector) can be shifted to non-negative numbers by adding a global offset to all the elements.

arXiv: 2104.13467

Setup Schematic



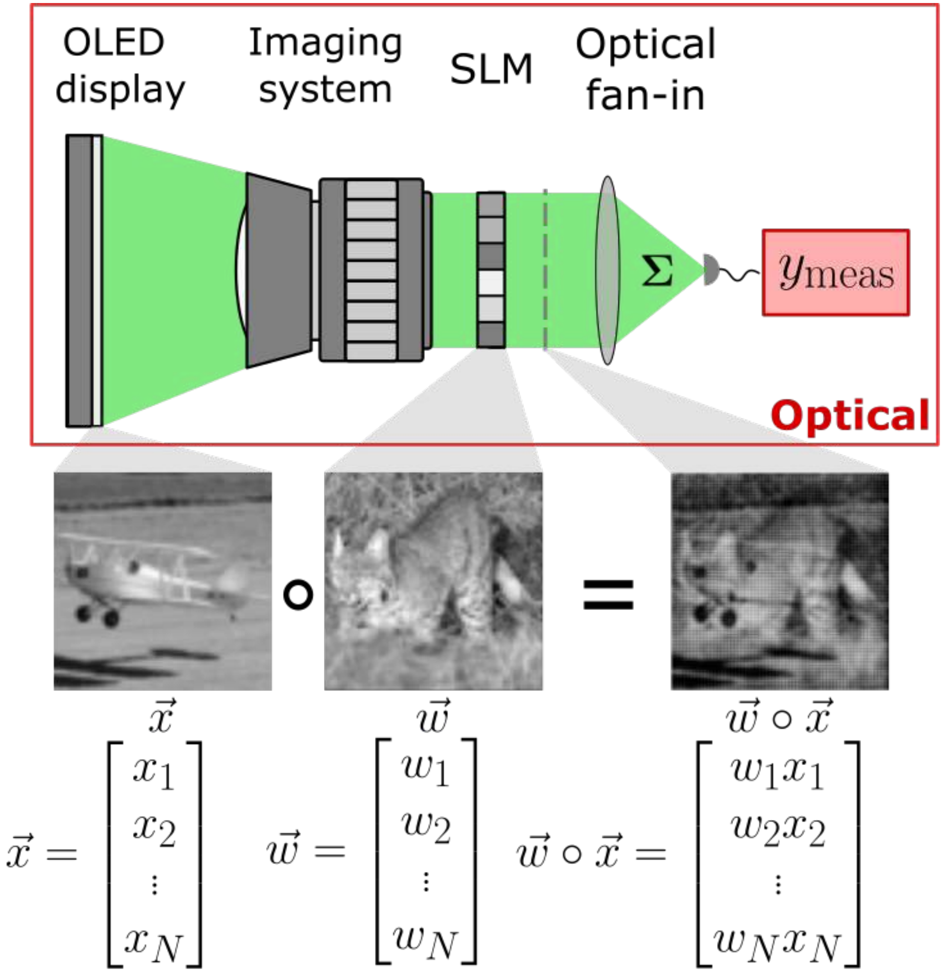
~0.5 million OLED pixels were aligned one-to-one to SLM pixels, which means the largest possible vector size in a vector-vector dot product was ~0.5 million.



3D modeling credit: Hannah Doyle

Numerical Accuracy of Vector Dot Products

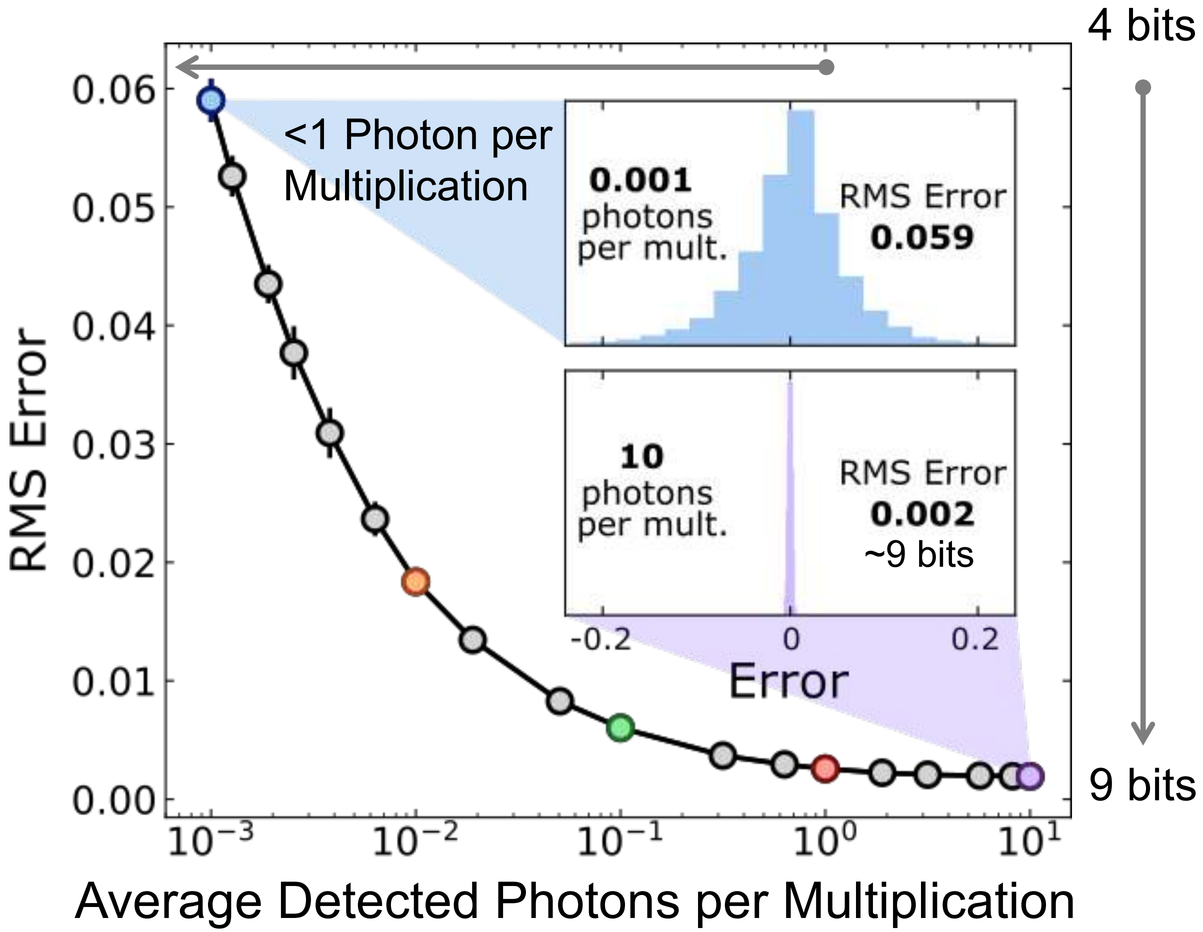
Dot products between random vectors (2D images)



$\text{Error} = y_{\text{meas}} - y_{\text{truth}}$

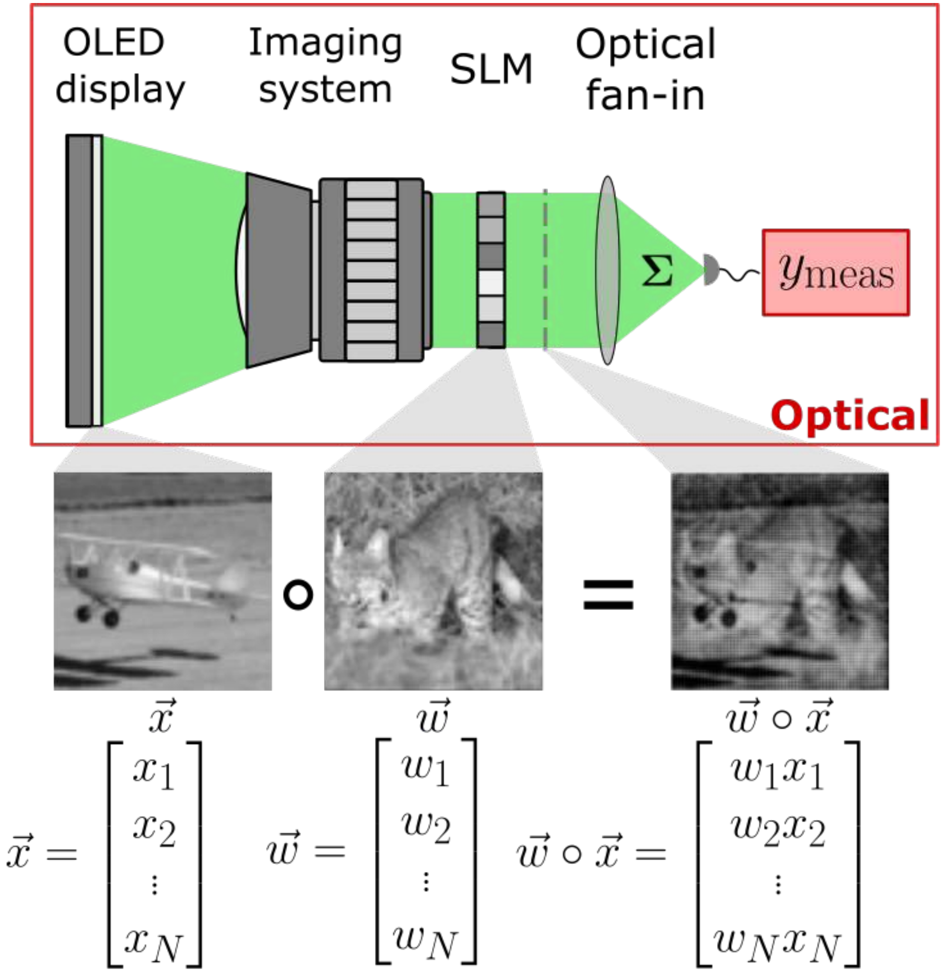
$y_{\text{meas}}, y_{\text{truth}}$ were normalized such that $y_{\text{truth}} \in [0, 1]$

Precision of vector-vector dot products for vector size $N \sim 500{,}000$



Numerical Accuracy of Vector Dot Products

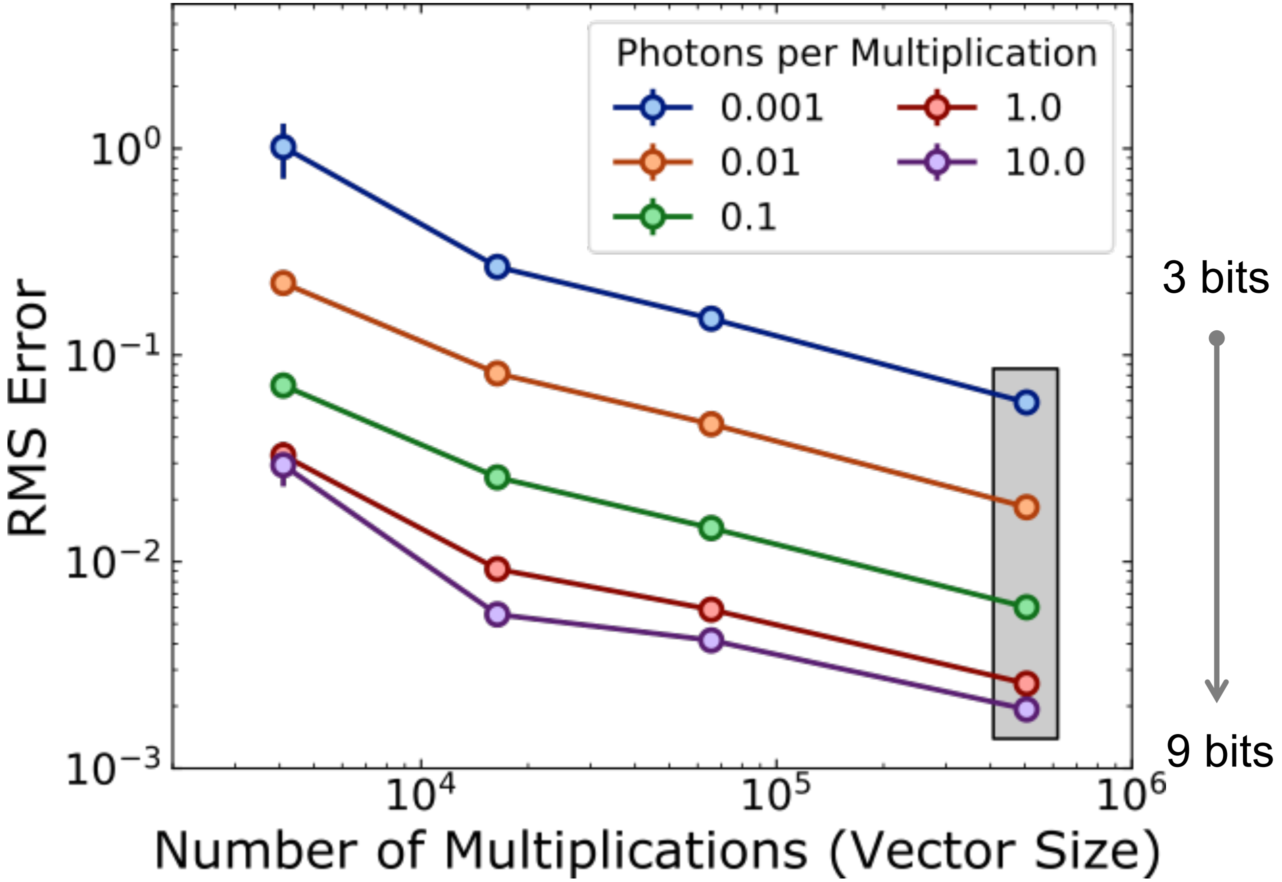
Dot products between random vectors (2D images)



$\text{Error} = y_{\text{meas}} - y_{\text{truth}}$

$y_{\text{meas}}, y_{\text{truth}}$ were normalized such that $y_{\text{truth}} \in [0, 1]$

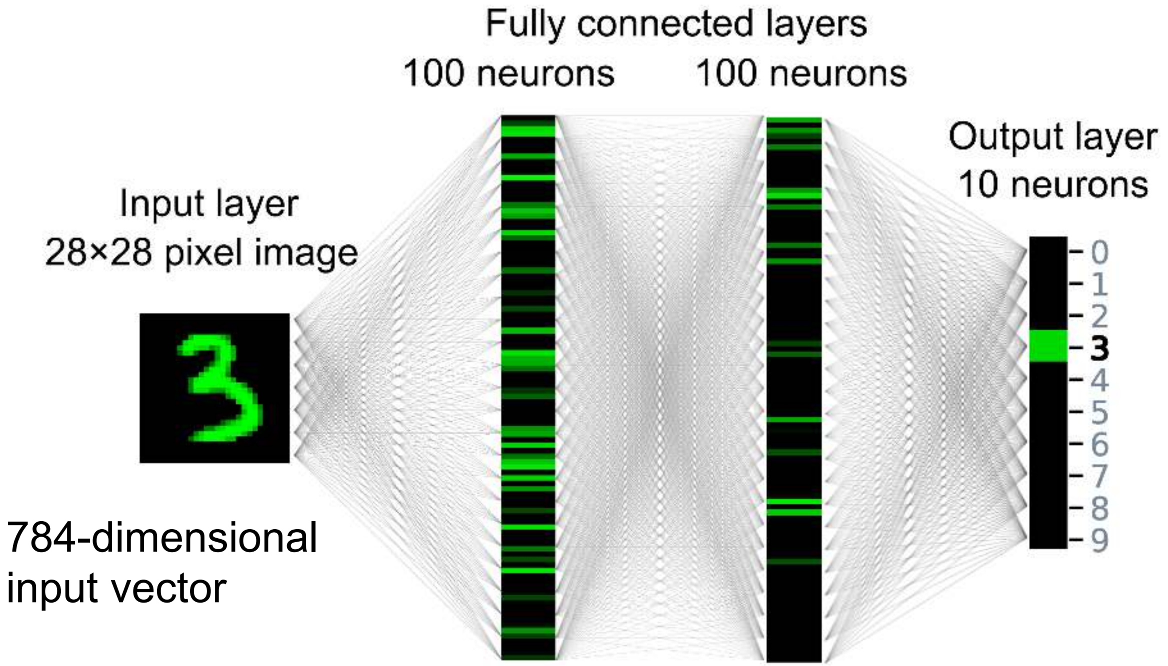
Precision of vector-vector dot products for different sizes and photon budgets



arXiv: 2104.13467

Experimental Results on Neural Networks

Given the good numerical accuracy in the sub-photon regime in the dot product test, can the ONN faithfully run a trained digital neural network?



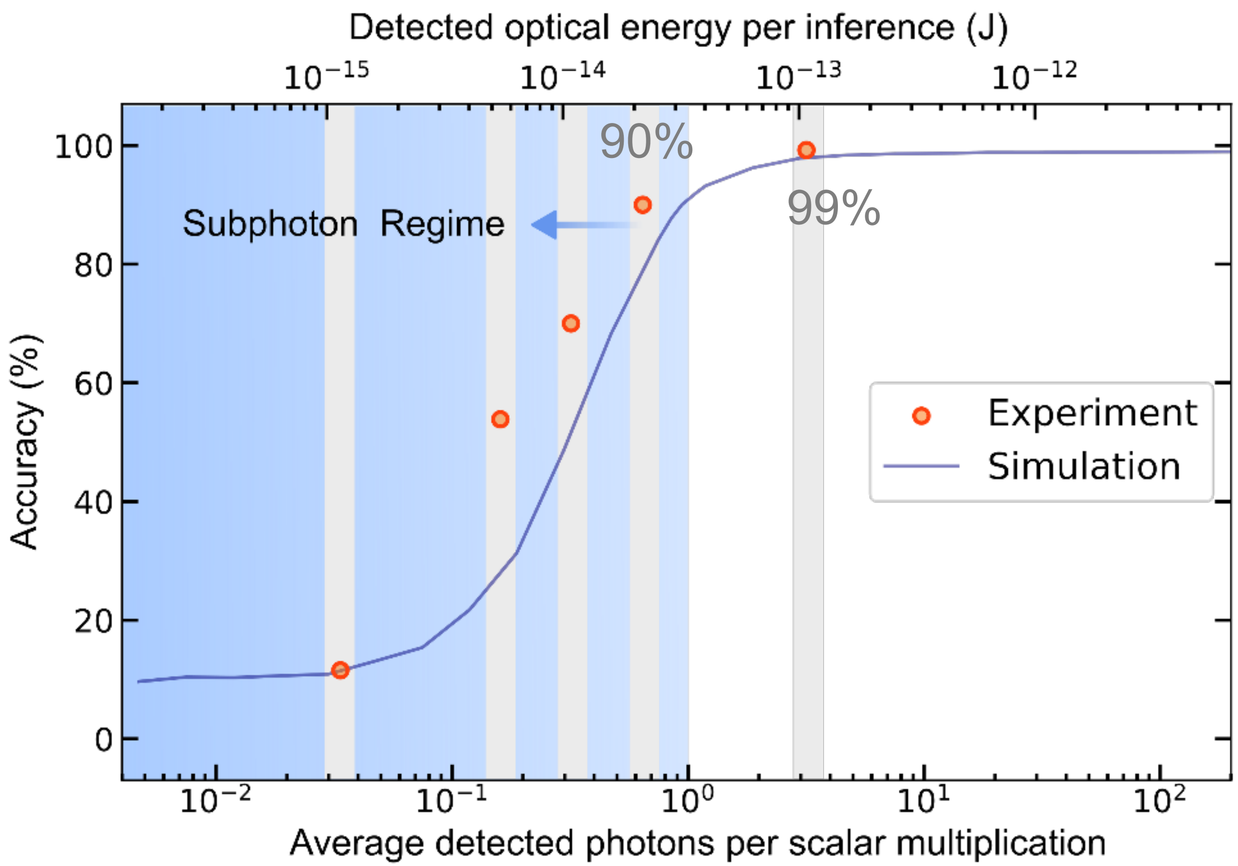
We used a multi-layer perceptron (784-100-100-10), trained with quantization aware training* to match with the hardware precision (~4 bits).

*B. Jacob *et al.* CVPR 2704-2713 (2017)

arXiv: 2104.13467

Classification Accuracy vs Photon Budget

High classification accuracy was obtained with even <1 photon per multiplication.



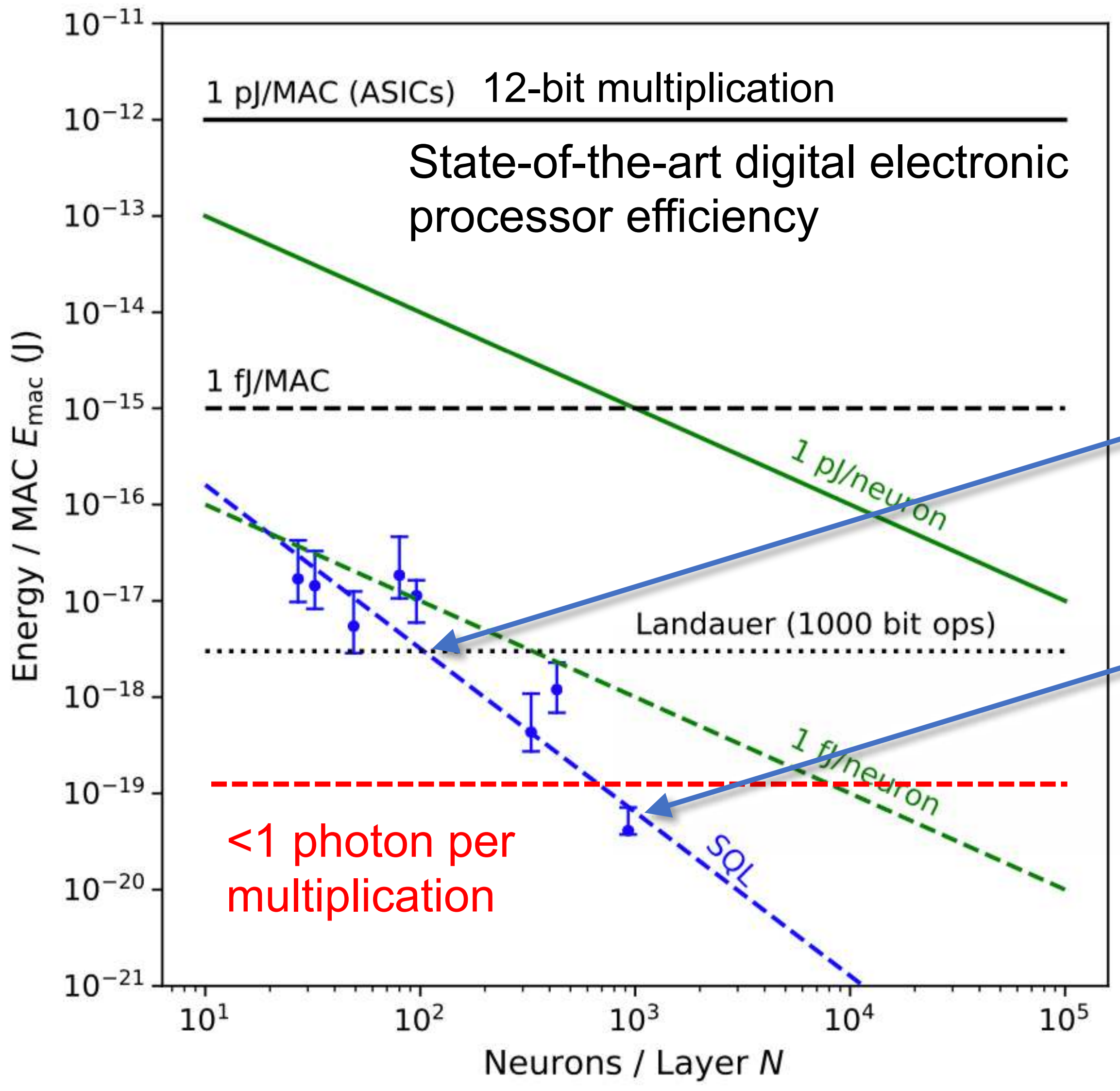
Experiment: Execute the ONN by performing *all* the MVM optically, with the controlled average detected photons per neuron at the output of each layer.

Simulation: Execute the NN model completely on a digital computer with simulated photon shot noise.

Note: the optical energy only refers to *detected* photons, which exclude optical loss or quantum efficiency of the detector.

* Average detected photons were scanned by changing detector integration time.

The Theoretical Limit of PNN Energy Consumption



The optical energy per scalar operation generally scales down with the vector size for PNNs*.

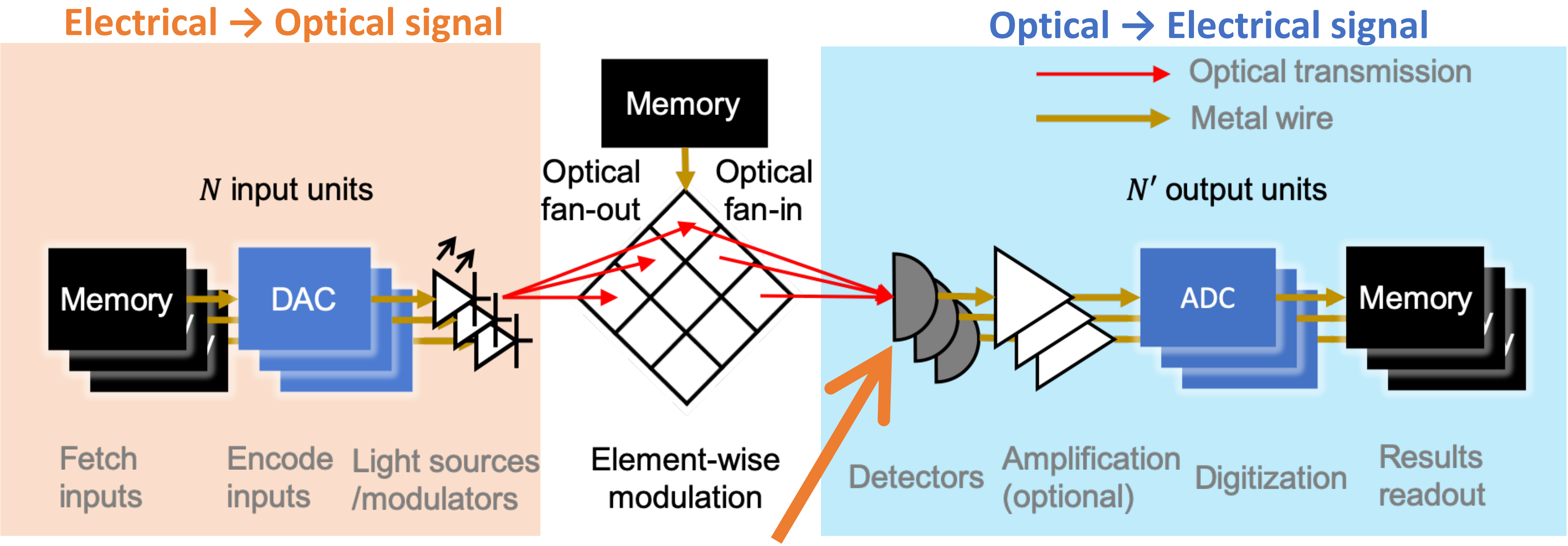
- Simulations indicate a vector size of 100-1000 is required for optics to achieve a lower optical energy consumption than the fundamental limit of digital computing.
- Less than 1 photon per scalar multiplication is possible with PNNs.
- Part of the energy efficiency stems from the robustness of PNNs to noise**, especially some loss of numerical precision is tolerable in a neural network.

* R. Hamerly, et al. *Phy. Rev.* X. **9,** 021032 (2019). (Figure source)
M. A. Nahmias, et al. *IEEE J. Sel. Top. Quantum Electron*. **26, 1** (2019).
A. N. Tait. *preprint on arXiv*: 2108.04819 (2021)

** N. Semenova, et al. *preprint on arXiv:2103.07413* (2021).

arXiv: 2104.13467

Detection vs Whole-system Energy Consumption



Our experiment showed the optical energy consumption here can be <1 photon per multiplication on average.

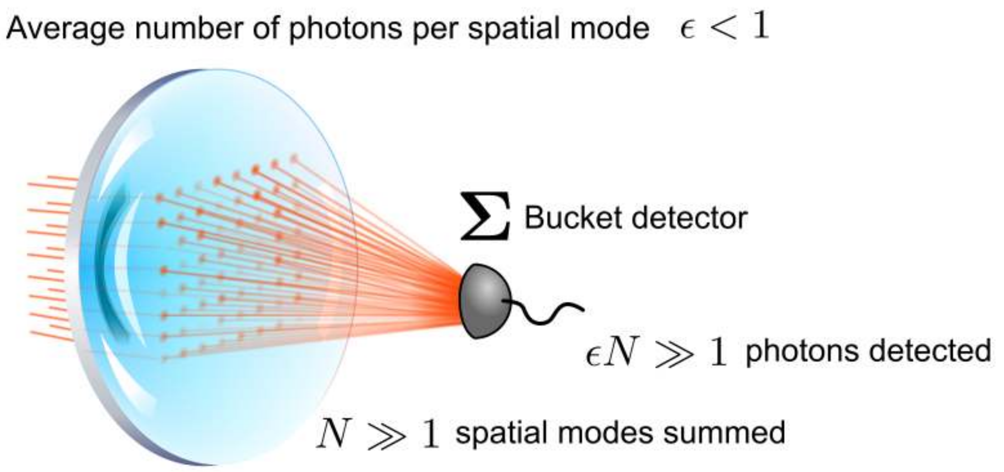
The whole-system energy consumption in an ideal system, including E → O → E conversions, can be similarly estimated as some other PNNs in the literature: 0.1-1 fJ/MAC for 3-5 bit resolution with sufficient scaling.

For details, see Section 15 of Supplementary Materials of *arXiv:* 2104.13467v1 (2021)

More resources: M. A. Nahmias, et al. *IEEE J. Sel. Top. Quantum Electron*. **26, 1** (2019).
A. N. Tait. *preprint on arXiv*: 2108.04819 (2021)

Conclusions

- We provide experiment evidence that photonic matrix-vector multiplication can achieve <1 detected photon per multiplication, even in a relatively small fully-connected neural network during machine learning inference.
- Our results support the estimation that photonic neural networks have the near-term potential to achieve an overall energy advantage over digital electronic processors.



T. Wang, S-Y, Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon.
An optical neural network using less than 1 photon per multiplication. *Preprint on arXiv:* 2104.13467 (2021)

GitHub: https://github.com/mcmahon-lab/ONN-QAT-SQL.git. (PNN training)
https://github.com/mcmahon-lab/ONN-device-control.git (PNN device control)

Email: tw329@cornell.edu, plm99@cornell.edu

Twitter: peterlmcmahon, TianYuWang2020