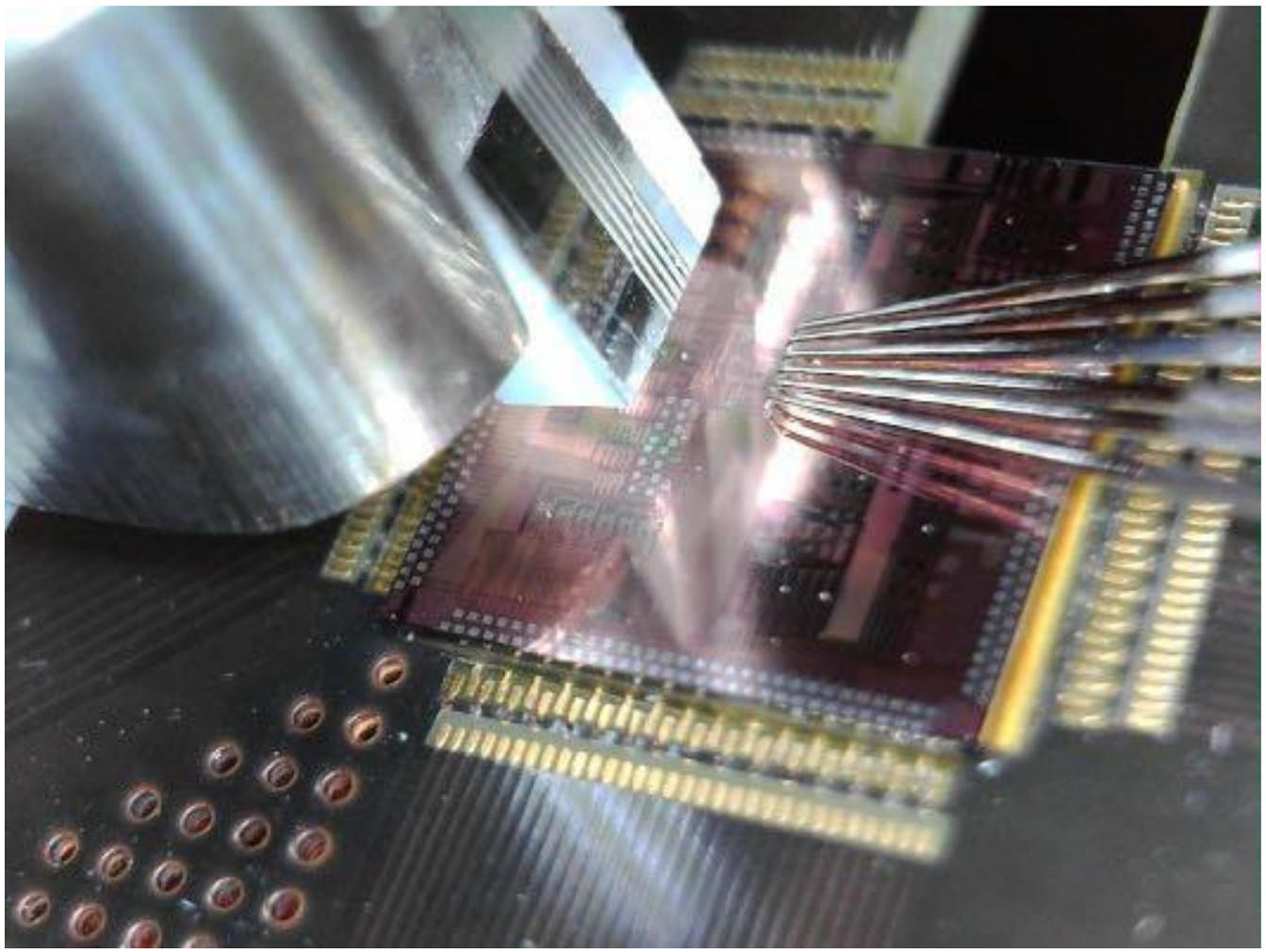
Single chip photonic deep neural network with accelerated training

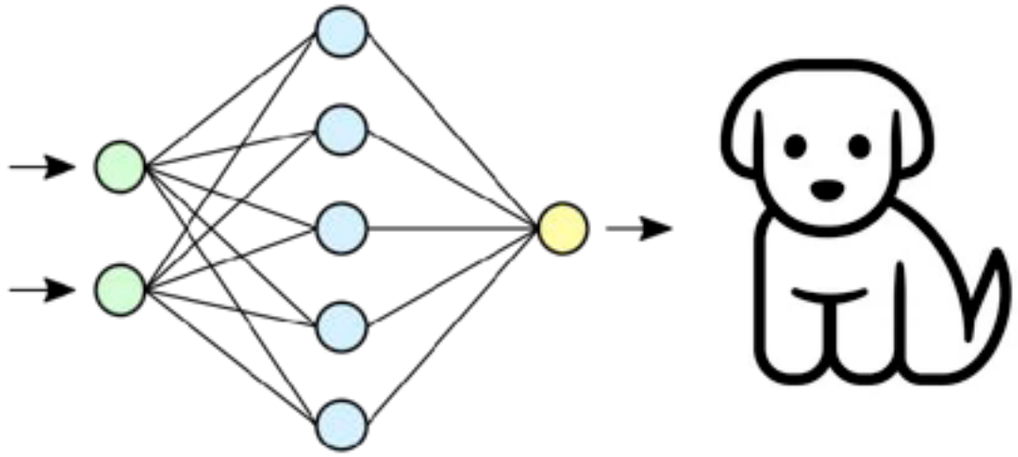
Saumil Bandyopadhyay[1], Alexander Sludds[1], Stefan Krastanov[1], Ryan Hamerly[1,2], Nicholas Harris[1], Darius Bunandar[1], Matthew Streshinsky[3], Michael Hochberg[4], Dirk Englund[1]

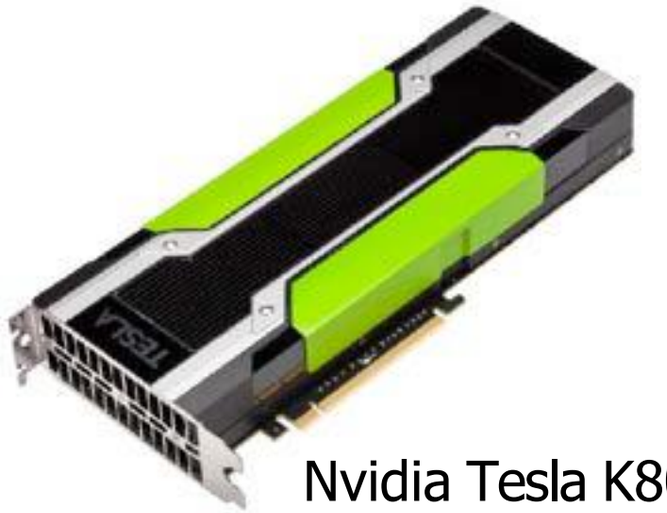
[1]Massachusetts Institute of Technology
[2]PHI Labs, NTT Research
[3]Nokia Corporation
[4]Luminous Computing



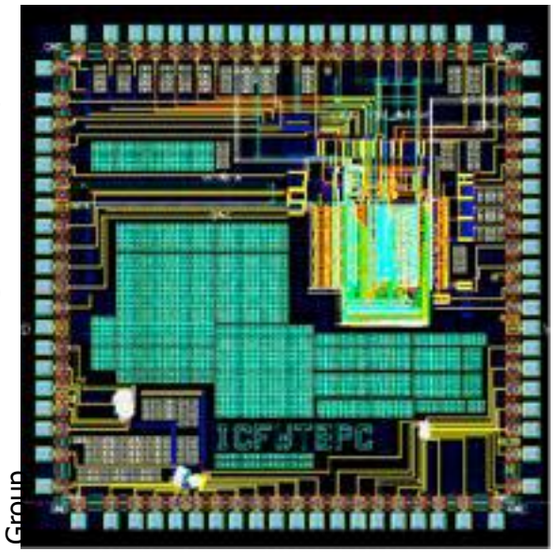
Preprint manuscript available at https://arxiv.org/abs/2208.01623

The “Cambrian explosion” of DNN hardware

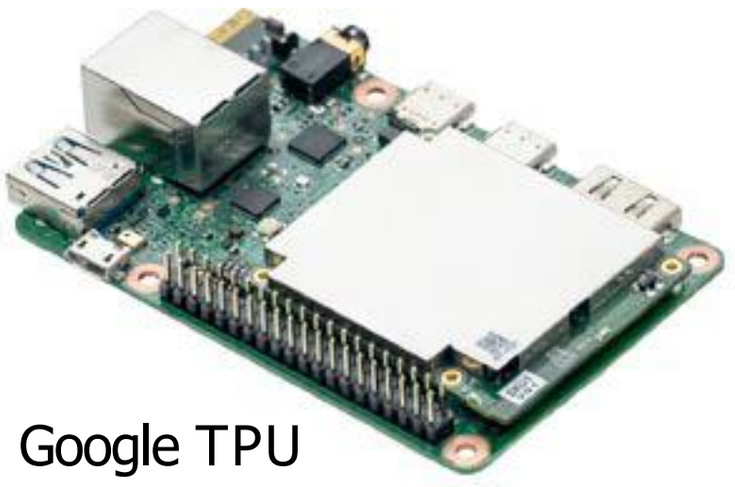




Nvidia Tesla K80



Waterloo CMOS Design and Reliability Group



Google TPU

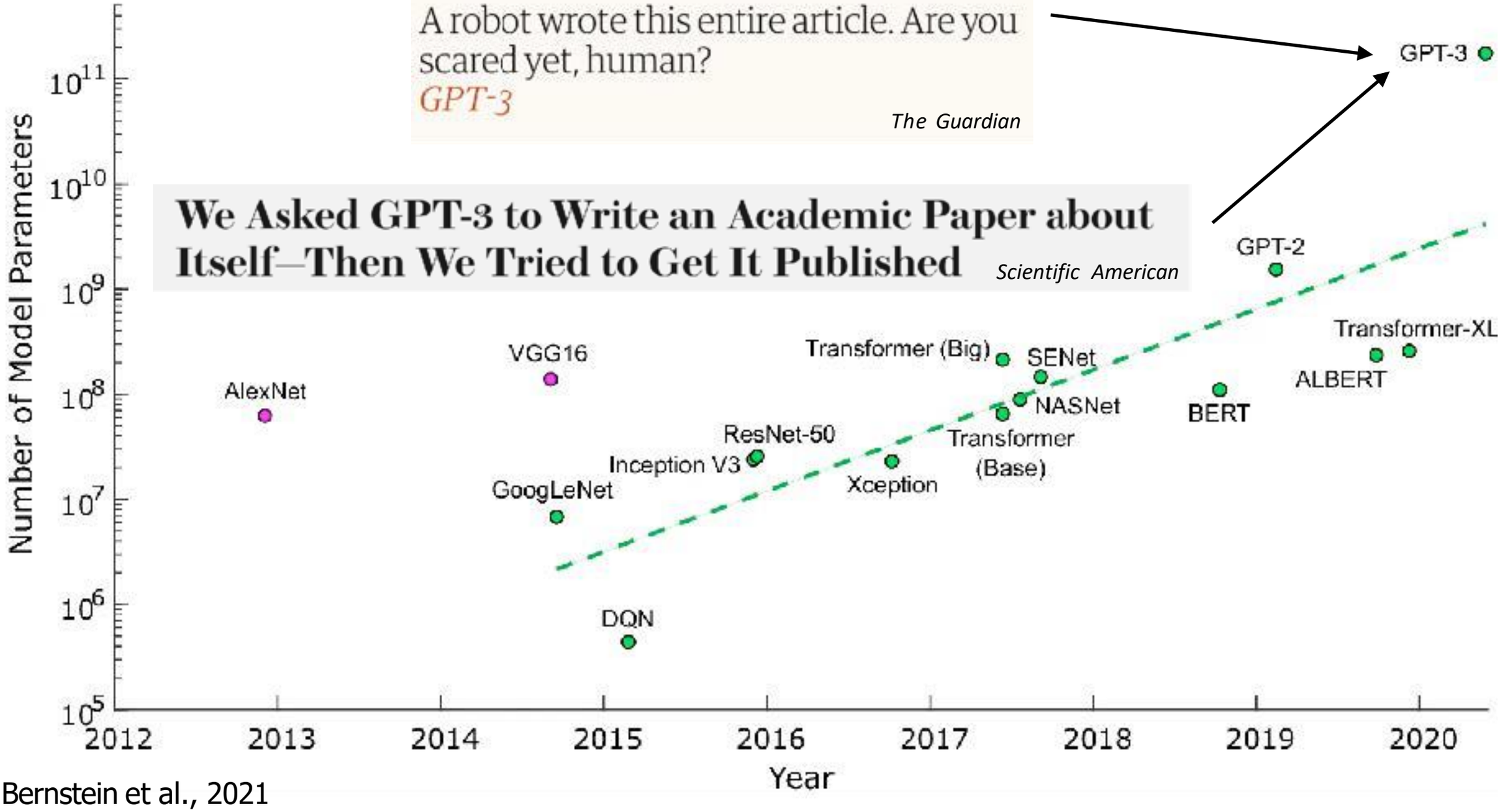
Preprint manuscript available at https://arxiv.org/abs/2208.01623

DNNs are getting better because they are getting *bigger*

A robot wrote this entire article. Are you scared yet, human?
GPT-3

The Guardian

We Asked GPT-3 to Write an Academic Paper about Itself—Then We Tried to Get It Published *Scientific American*

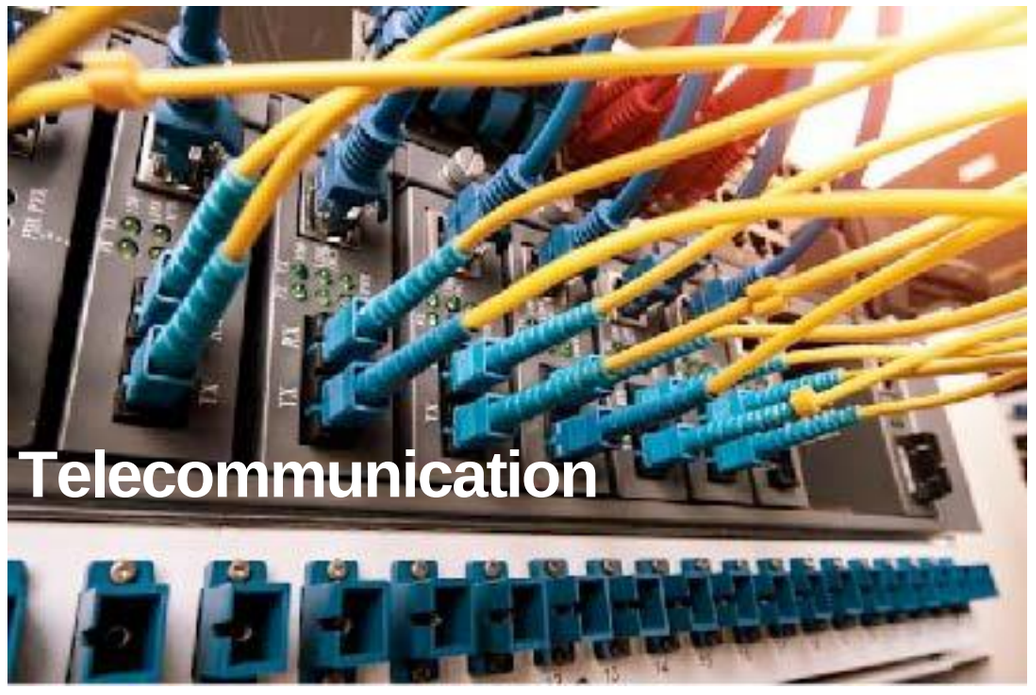


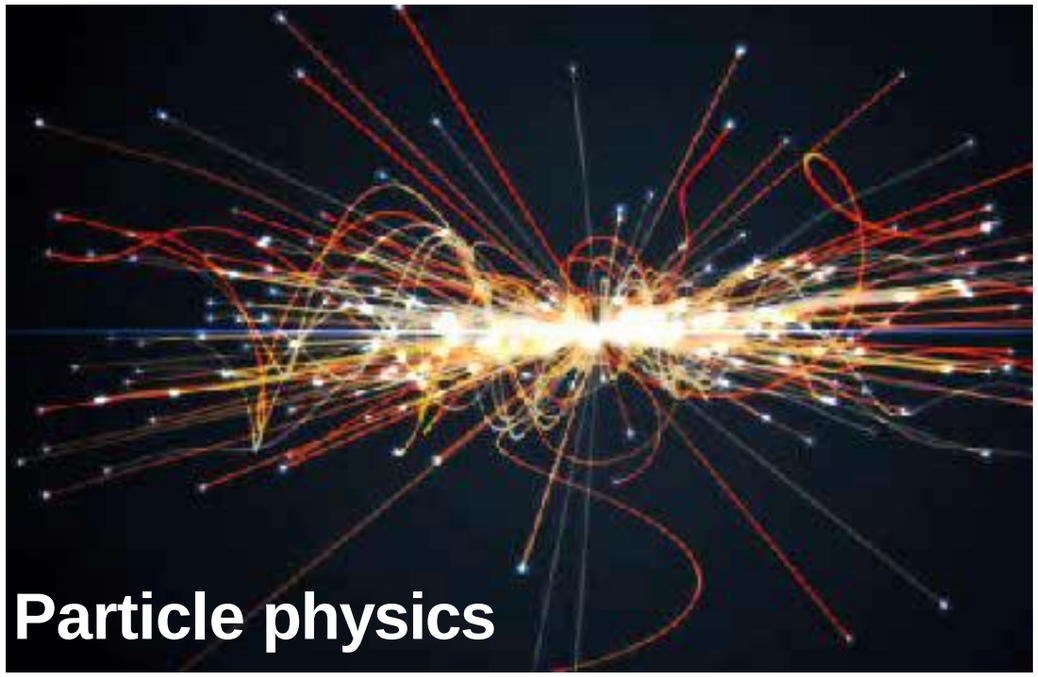
Bernstein et al., 2021

DNN scaling is ***bottlenecked*** by energy consumption, which is ~**1 pJ/OP** for CMOS.

Preprint manuscript available at https://arxiv.org/abs/2208.01623

New DNN hardware is needed for ultra-low latency processing

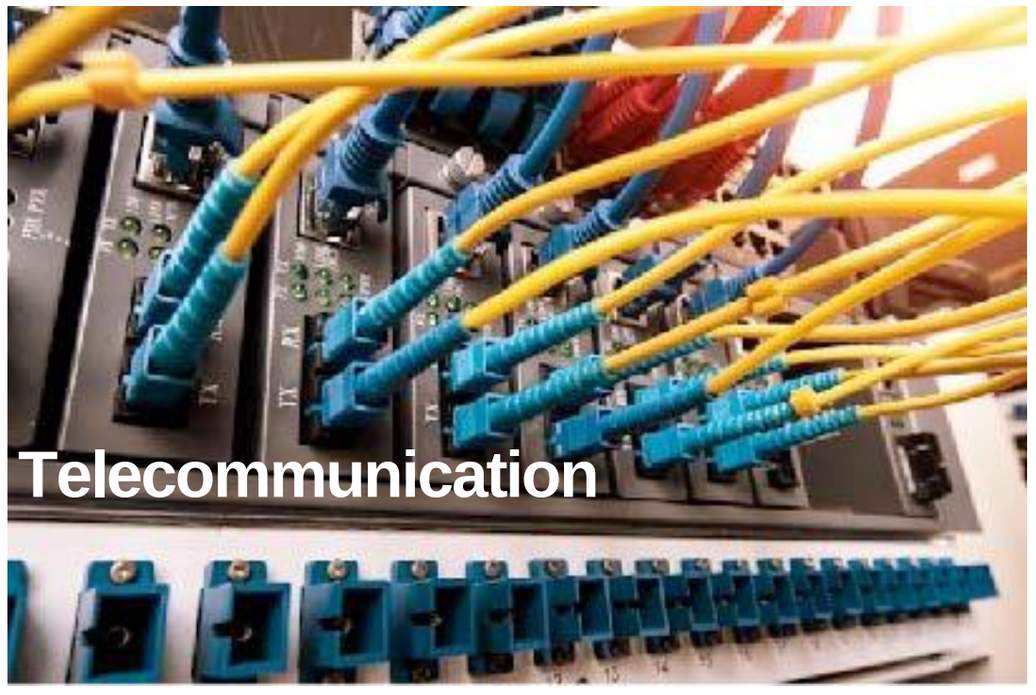


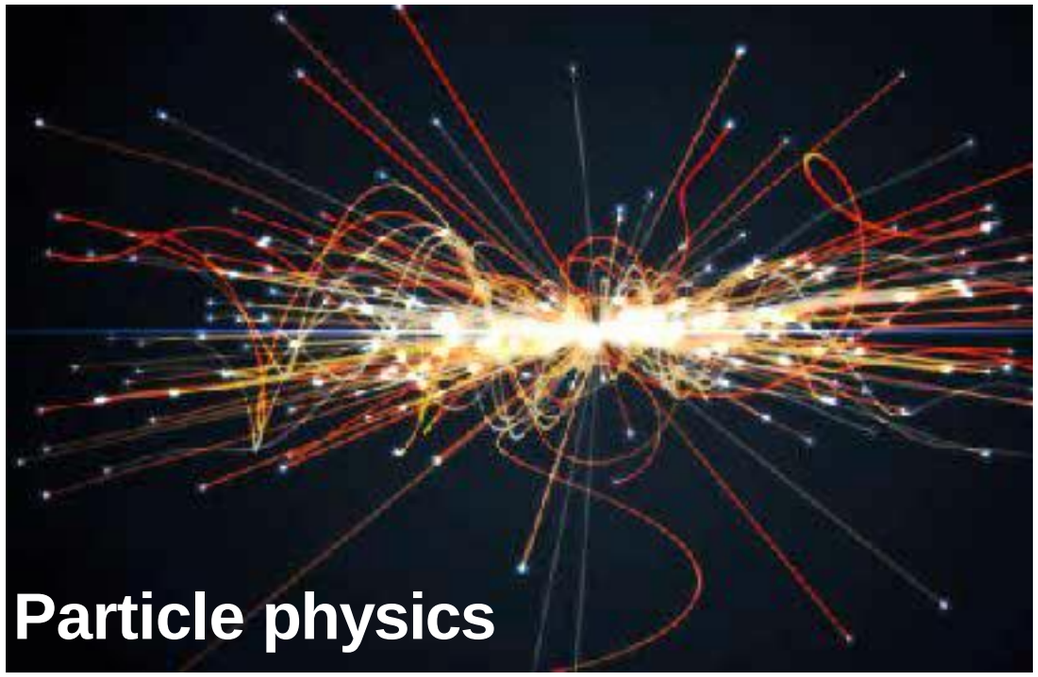


New applications require ultra-low latency (**µs-ns**) processing of (**optical**) data

Preprint manuscript available at https://arxiv.org/abs/2208.01623

New DNN hardware is needed for ultra-low latency processing



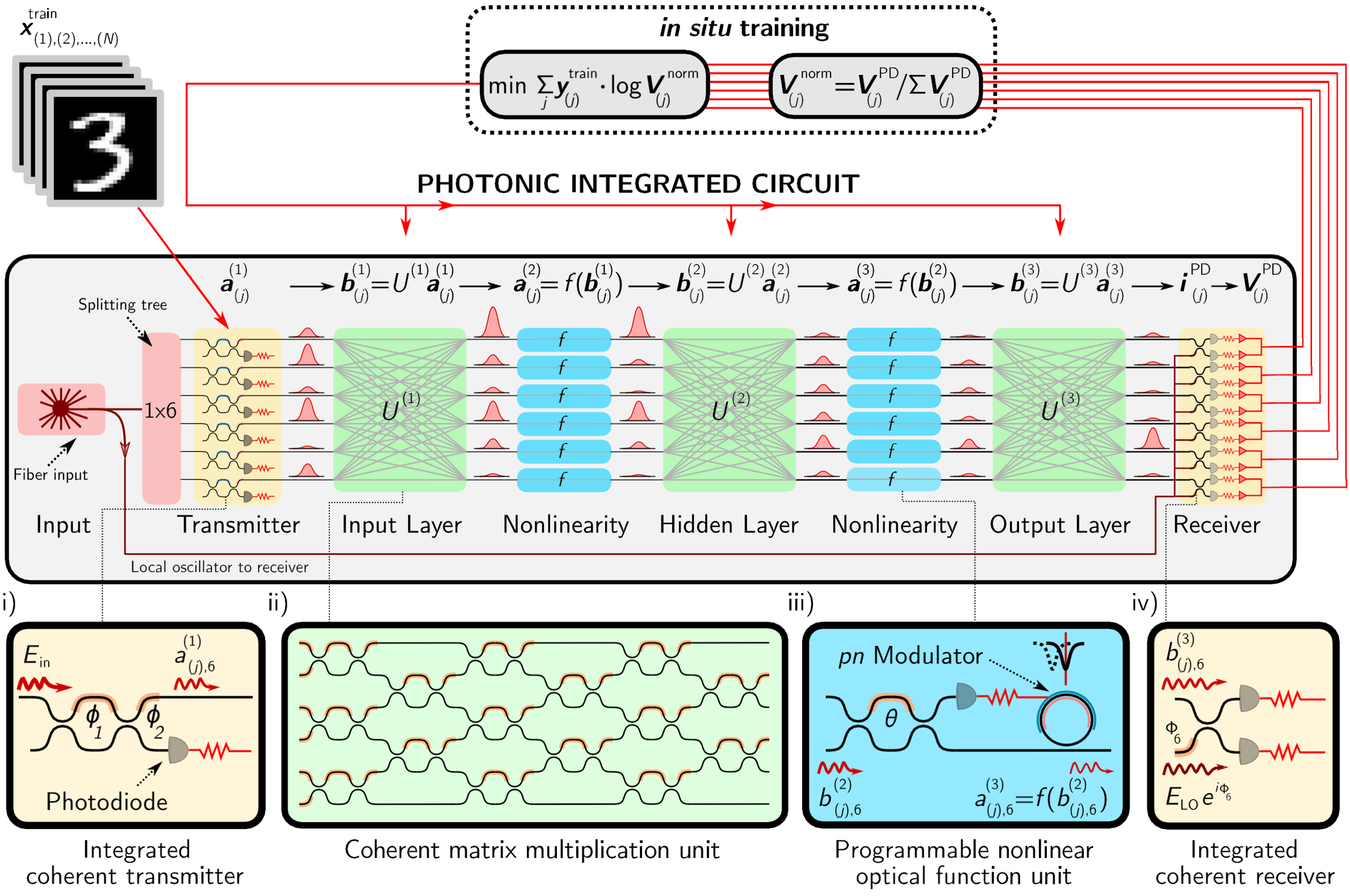


New applications require ultra-low latency (**µs-ns**) processing of (**optical**) data

This requires time-of-flight (**clockless**) compute on a **single, integrated chip**

Preprint manuscript available at https://arxiv.org/abs/2208.01623

An end-to-end optical DNN processor on a single chip



Integrated coherent transmitter

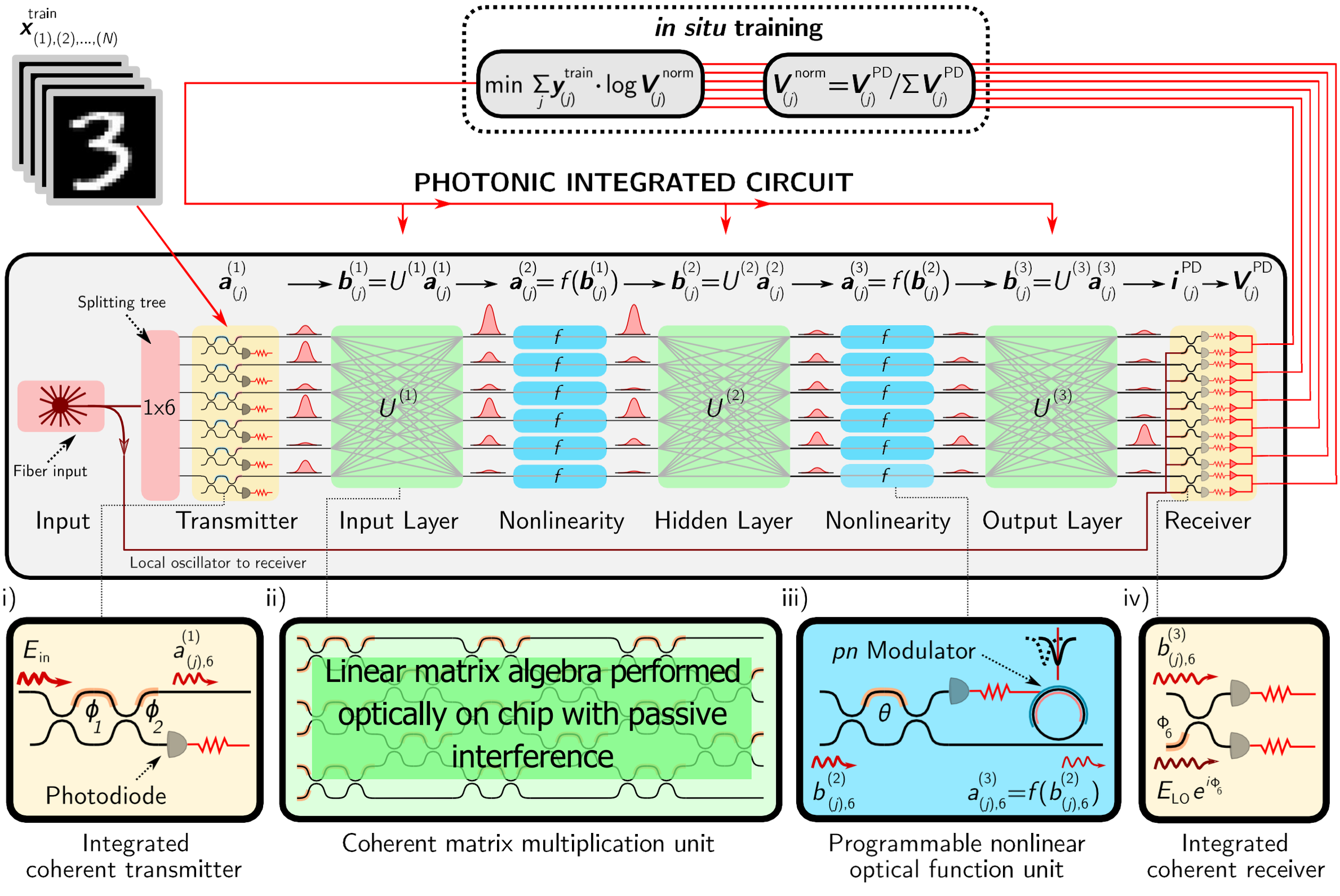
Coherent matrix multiplication unit

Programmable nonlinear optical function unit

Integrated coherent receiver

Preprint manuscript available at https://arxiv.org/abs/2208.01623

An end-to-end optical DNN processor on a single chip



Integrated coherent transmitter

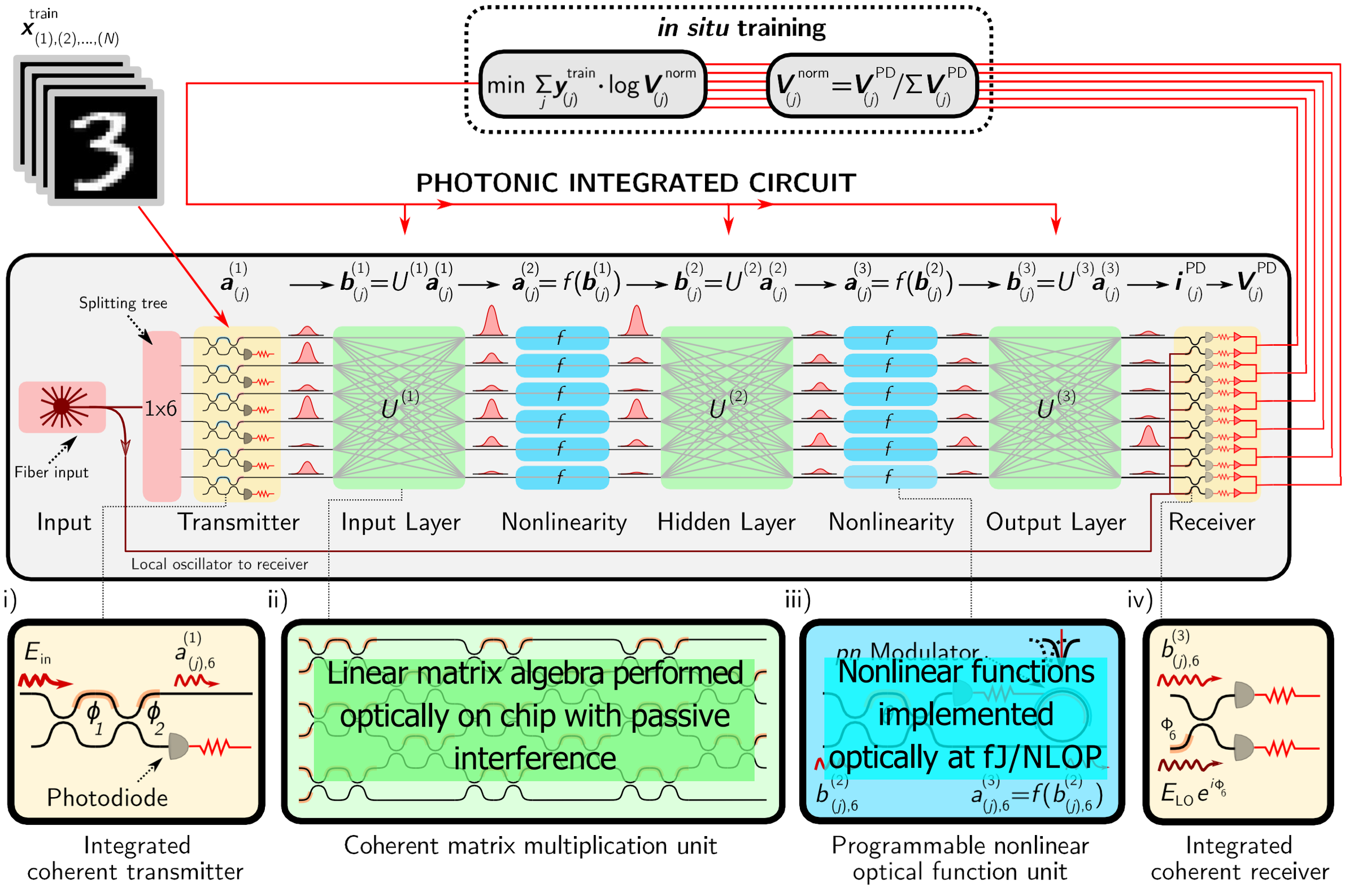
Coherent matrix multiplication unit

Programmable nonlinear optical function unit

Integrated coherent receiver

Preprint manuscript available at https://arxiv.org/abs/2208.01623

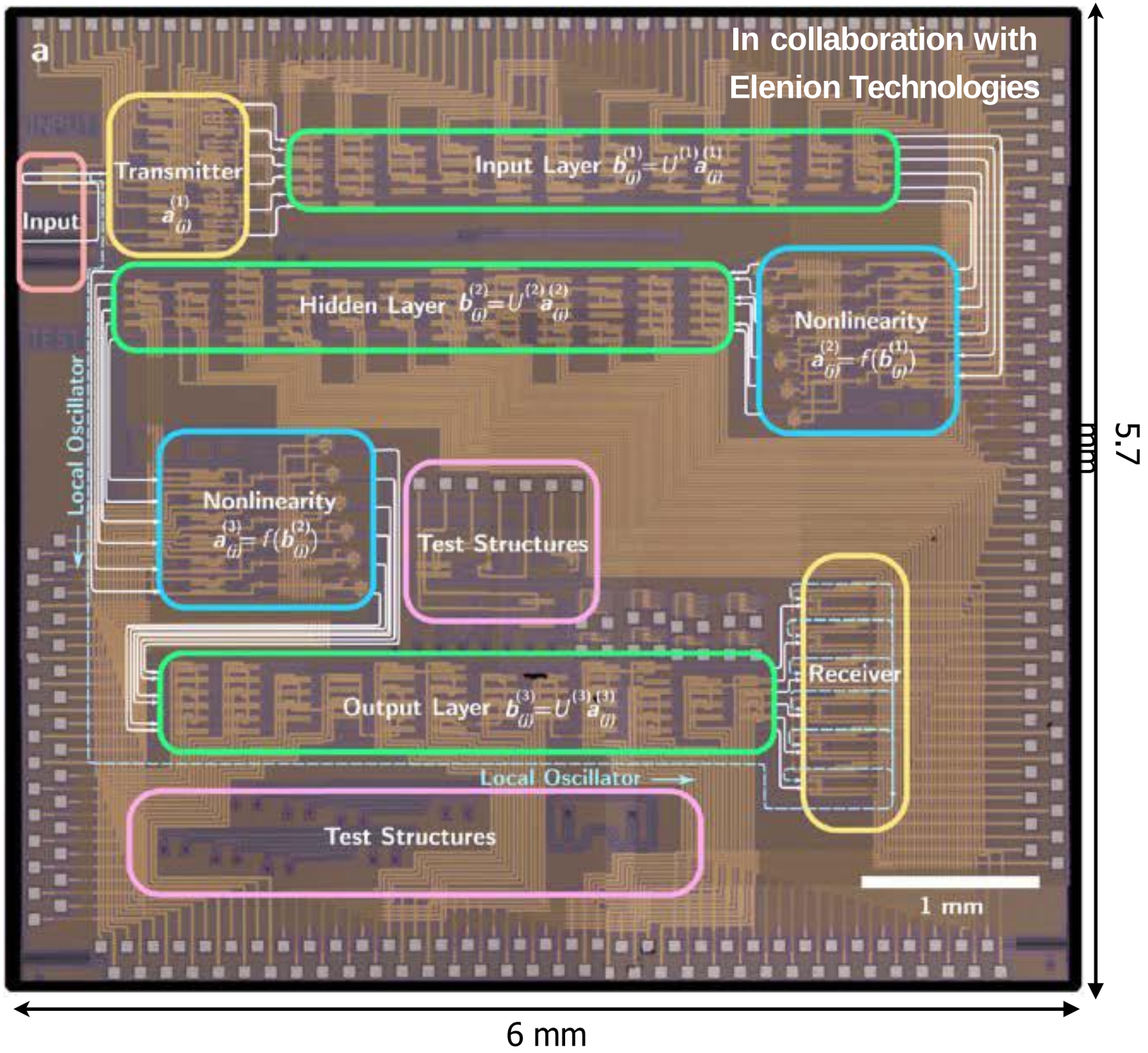
An end-to-end optical DNN processor on a single chip



Integrated coherent transmitter | Coherent matrix multiplication unit | Programmable nonlinear optical function unit | Integrated coherent receiver

Preprint manuscript available at https://arxiv.org/abs/2208.01623

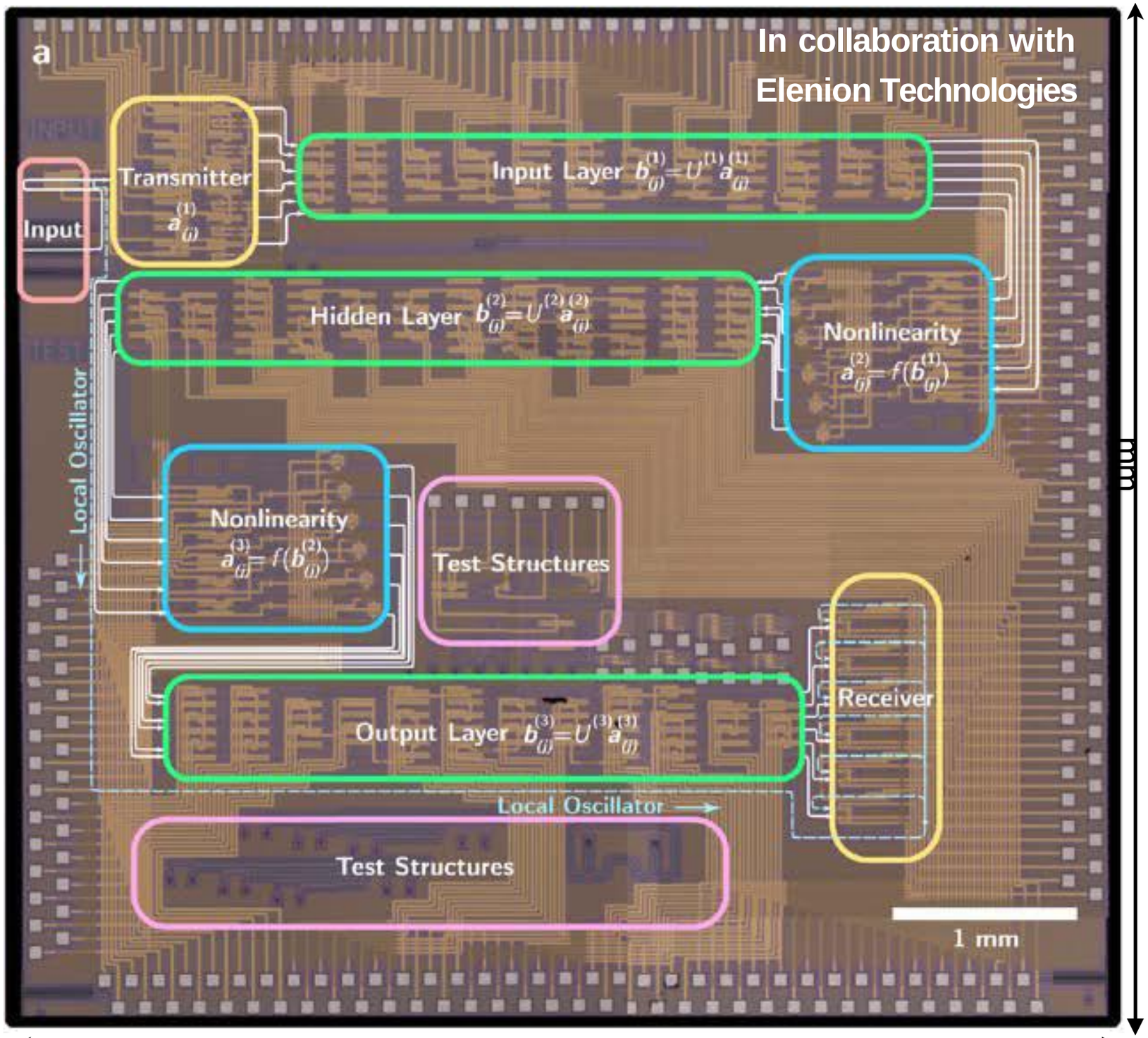
Application-specific photonic integrated circuit

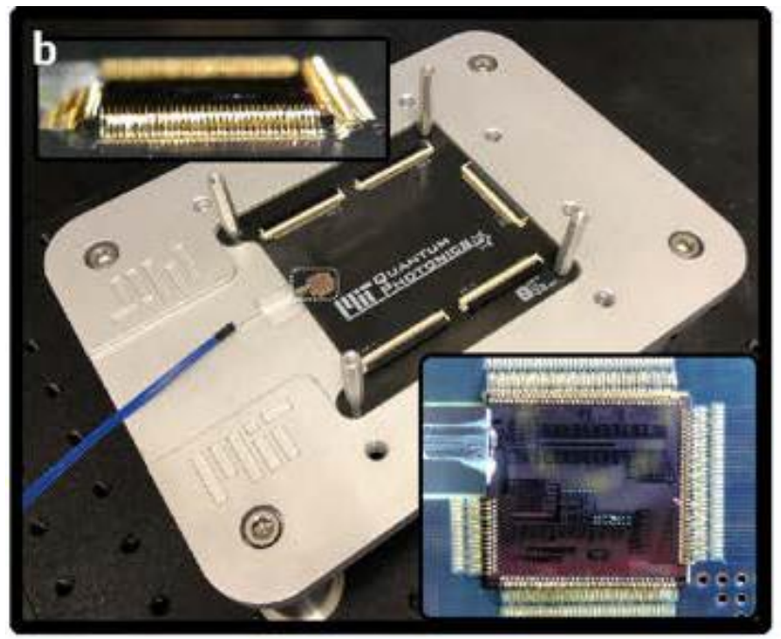


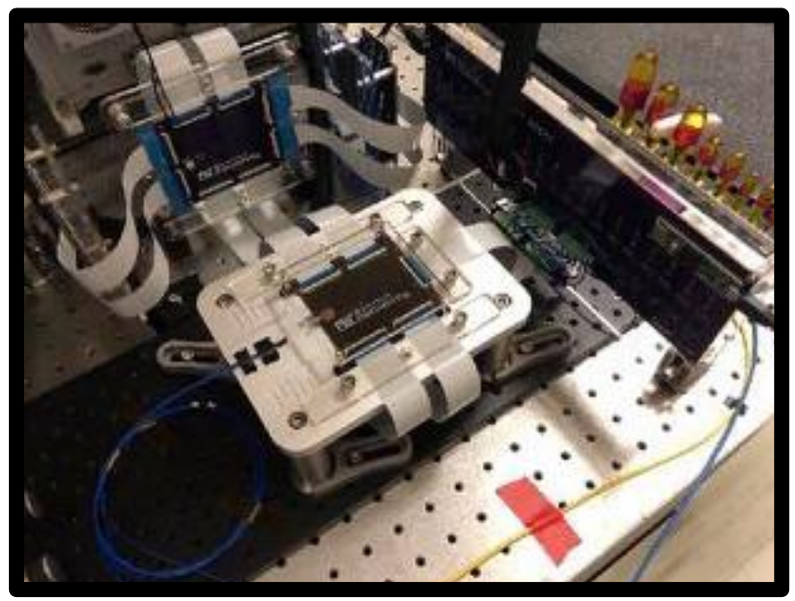
132 model parameters on-chip, 90 layers of optical components

Preprint manuscript available at https://arxiv.org/abs/2208.01623

Application-specific photonic integrated circuit



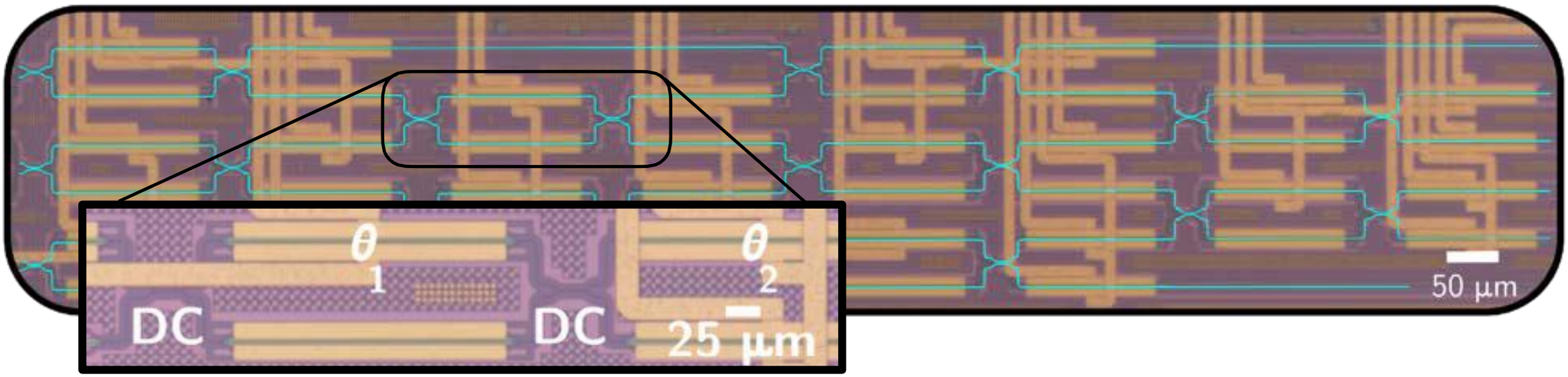


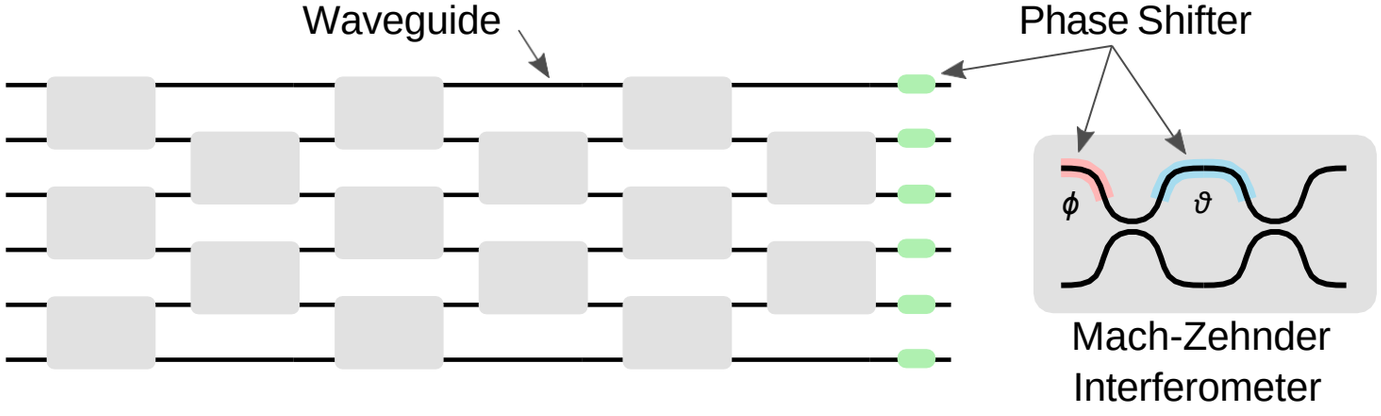


132 model parameters on-chip, 90 layers of optical components

Preprint manuscript available at https://arxiv.org/abs/2208.01623

Coherent matrix multiplication unit



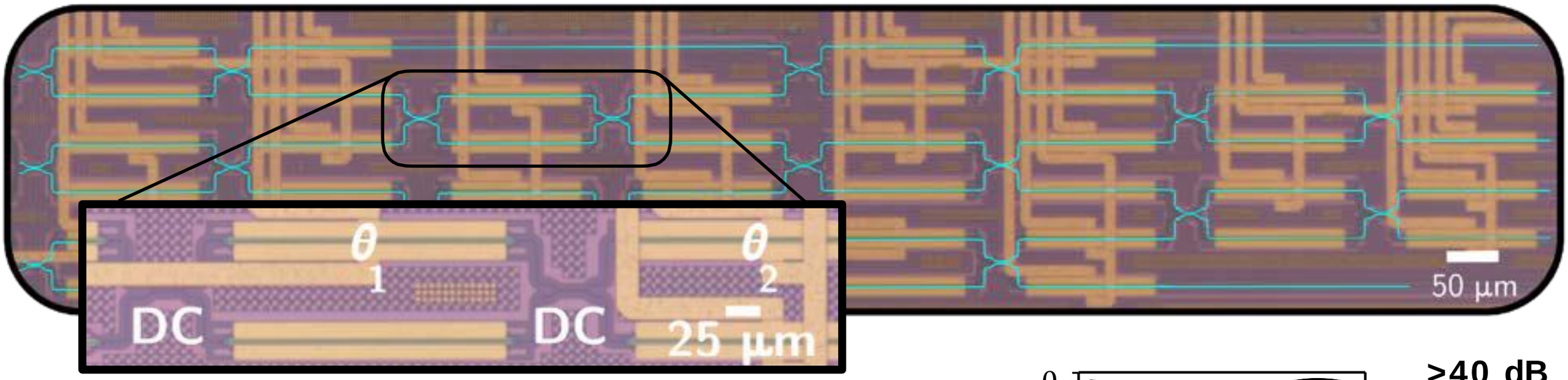


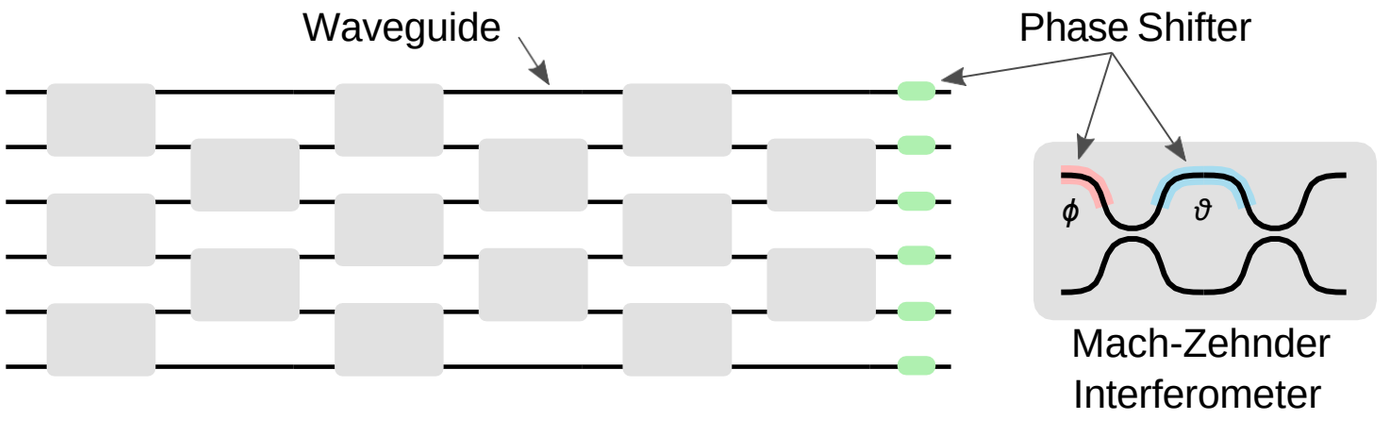
$$T_{ij}(\theta,\phi) = ie^{i\theta/2}\begin{bmatrix} e^{i\phi}\sin(\theta/2) & \cos(\theta/2) \\ e^{i\phi}\cos(\theta/2) & -\sin(\theta/2) \end{bmatrix} \qquad U = D\prod T_{ij}(\theta,\phi)$$

Programmable meshes of Mach-Zehnder interferometers compute linear algebra through passive optical interference

Preprint manuscript available at https://arxiv.org/abs/2208.01623

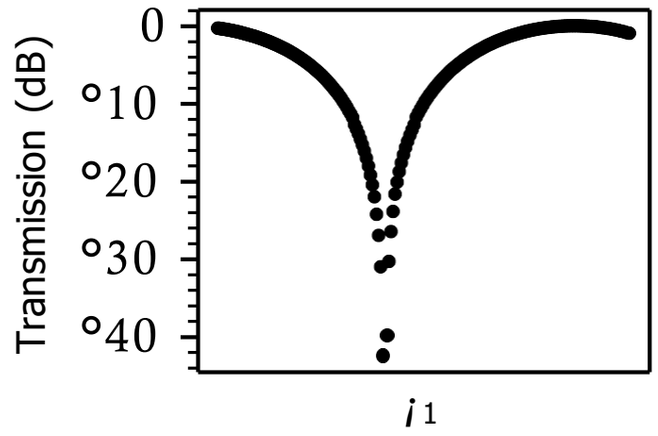
Coherent matrix multiplication unit



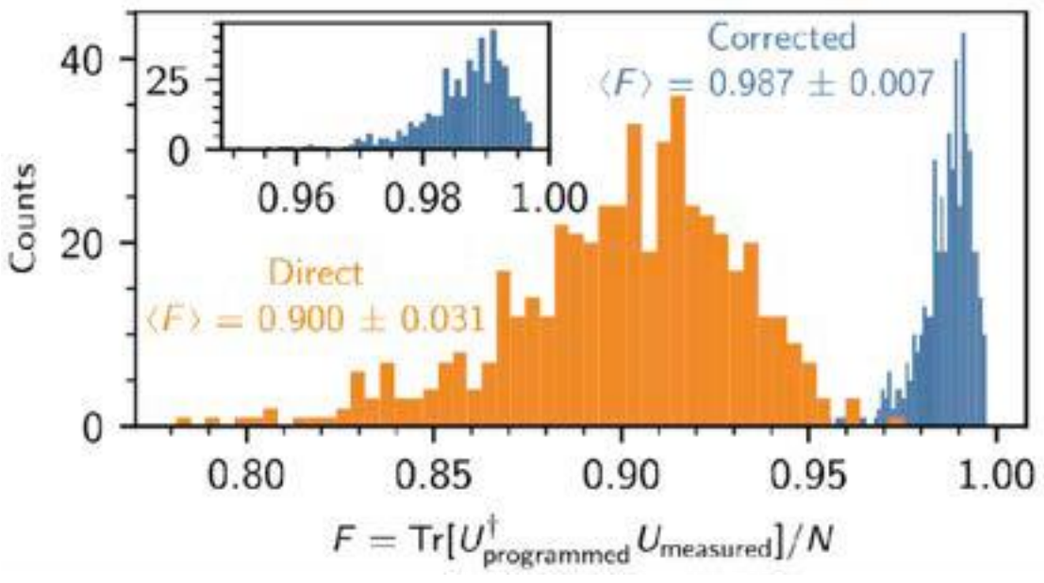


$$T_{ij}(\theta,\phi) = ie^{i\theta/2} \begin{bmatrix} e^{i\phi}\sin(\theta/2) & \cos(\theta/2) \\ e^{i\phi}\cos(\theta/2) & -\sin(\theta/2) \end{bmatrix} \qquad U = D\prod T_{ij}(\theta,\phi)$$

Programmable meshes of Mach-Zehnder interferometers compute linear algebra through passive optical interference

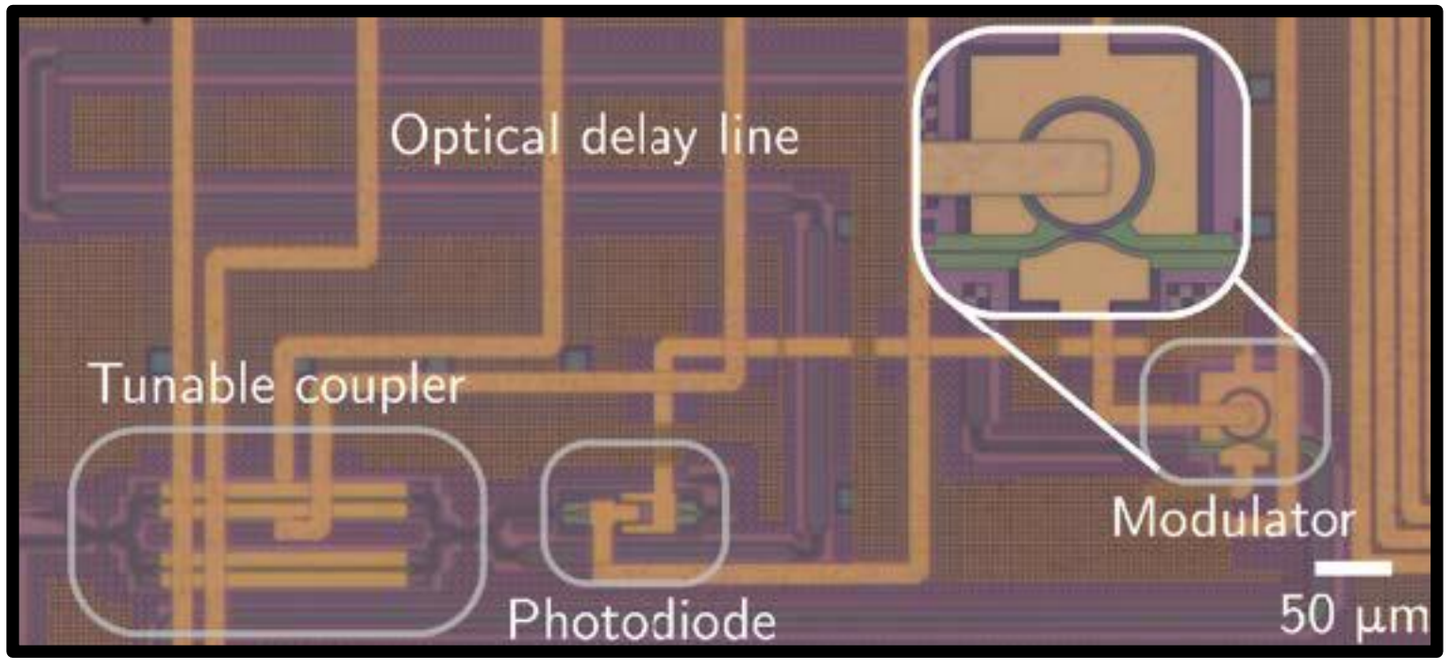


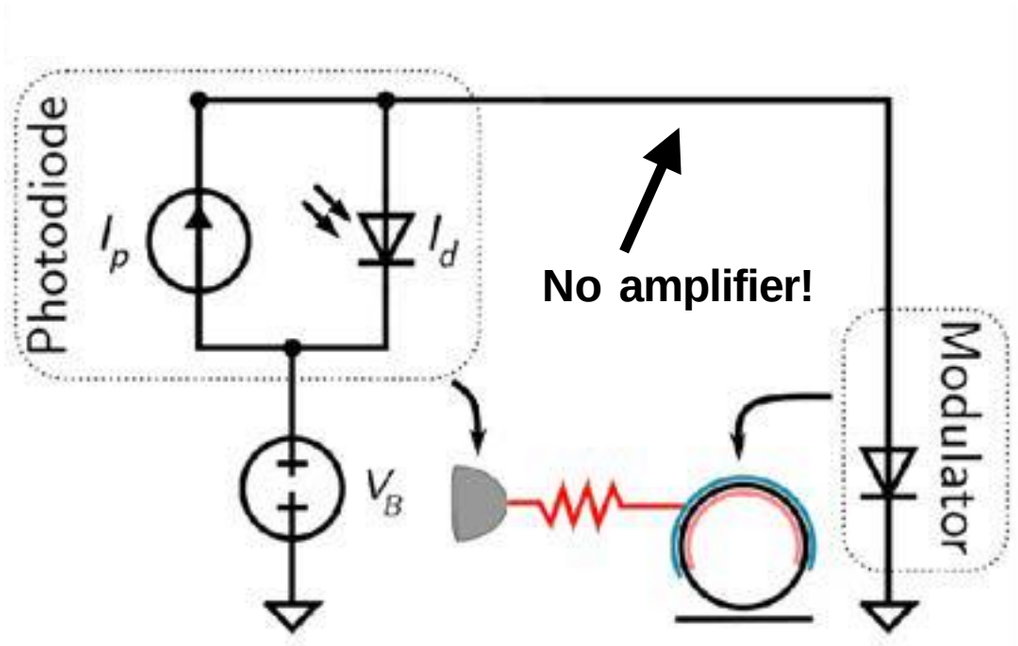
>40 dB optical extinction, >13 bits of precision on inputs



Preprint manuscript available at https://arxiv.org/abs/2208.01623

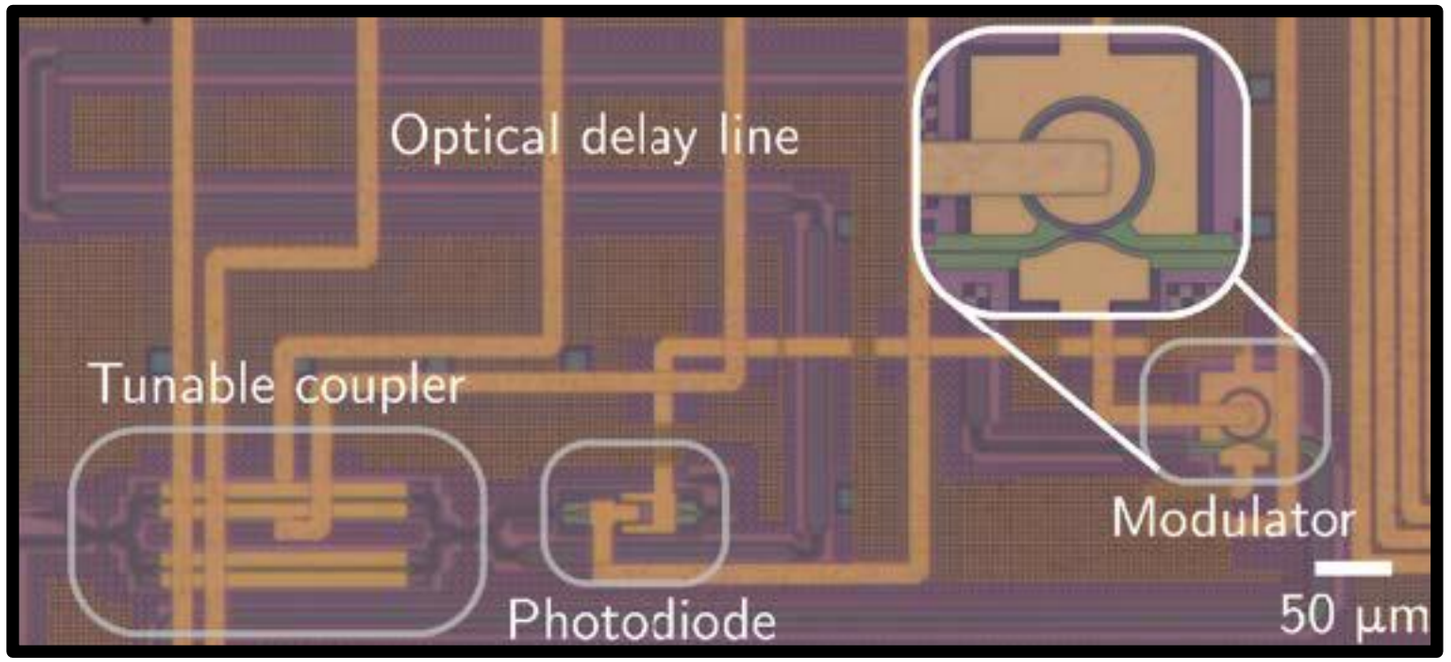
Nonlinear optical function unit

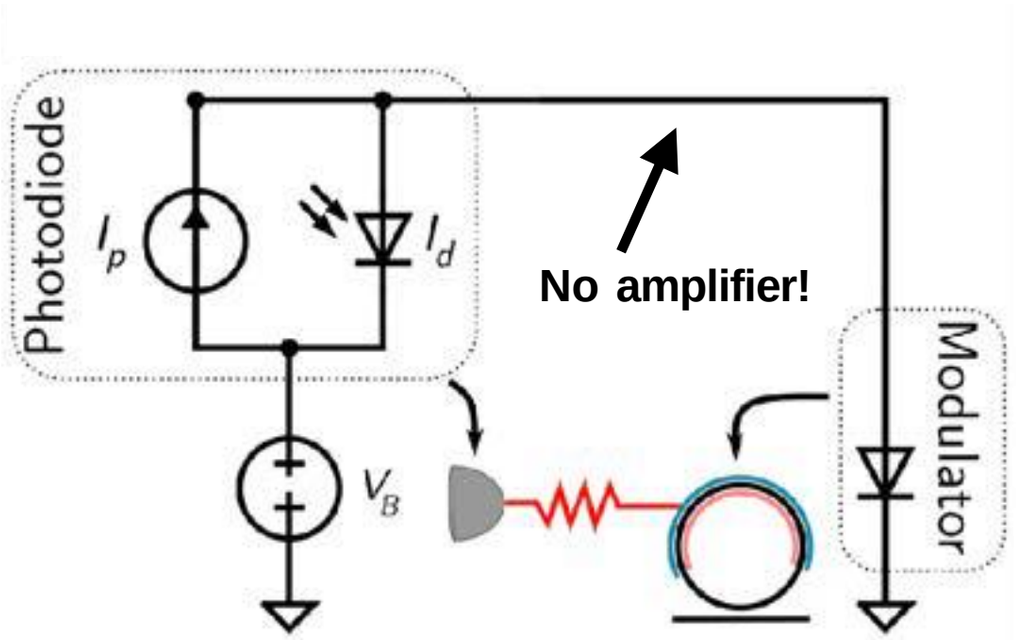


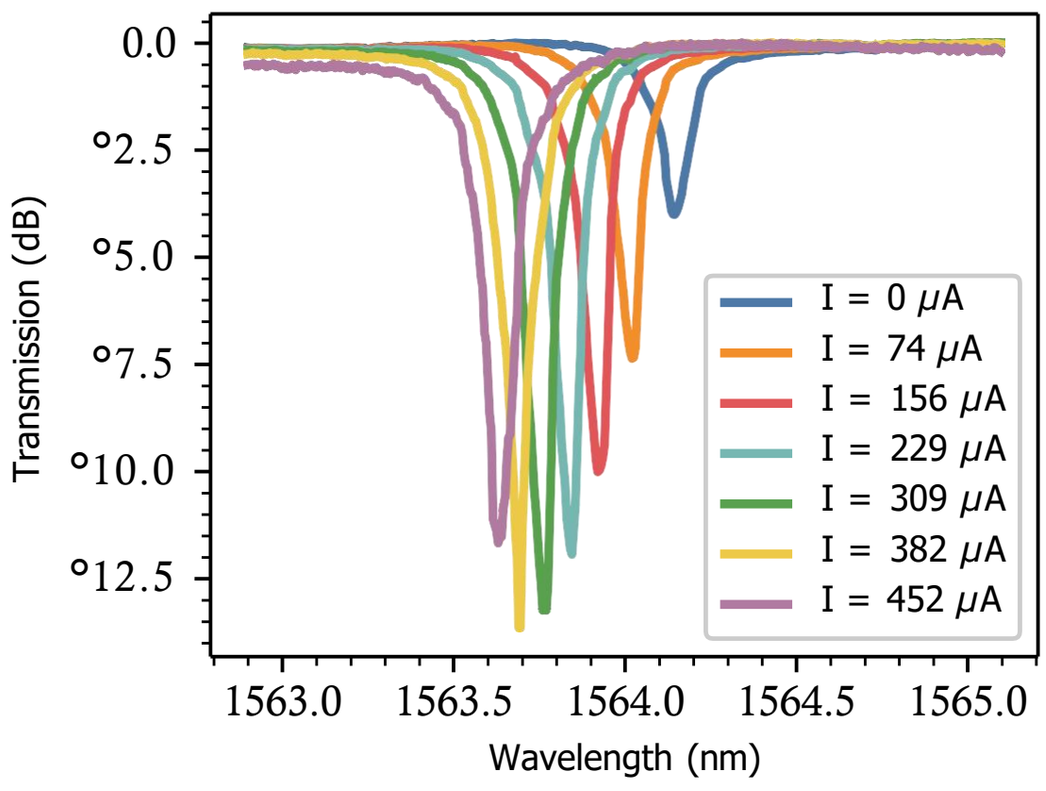


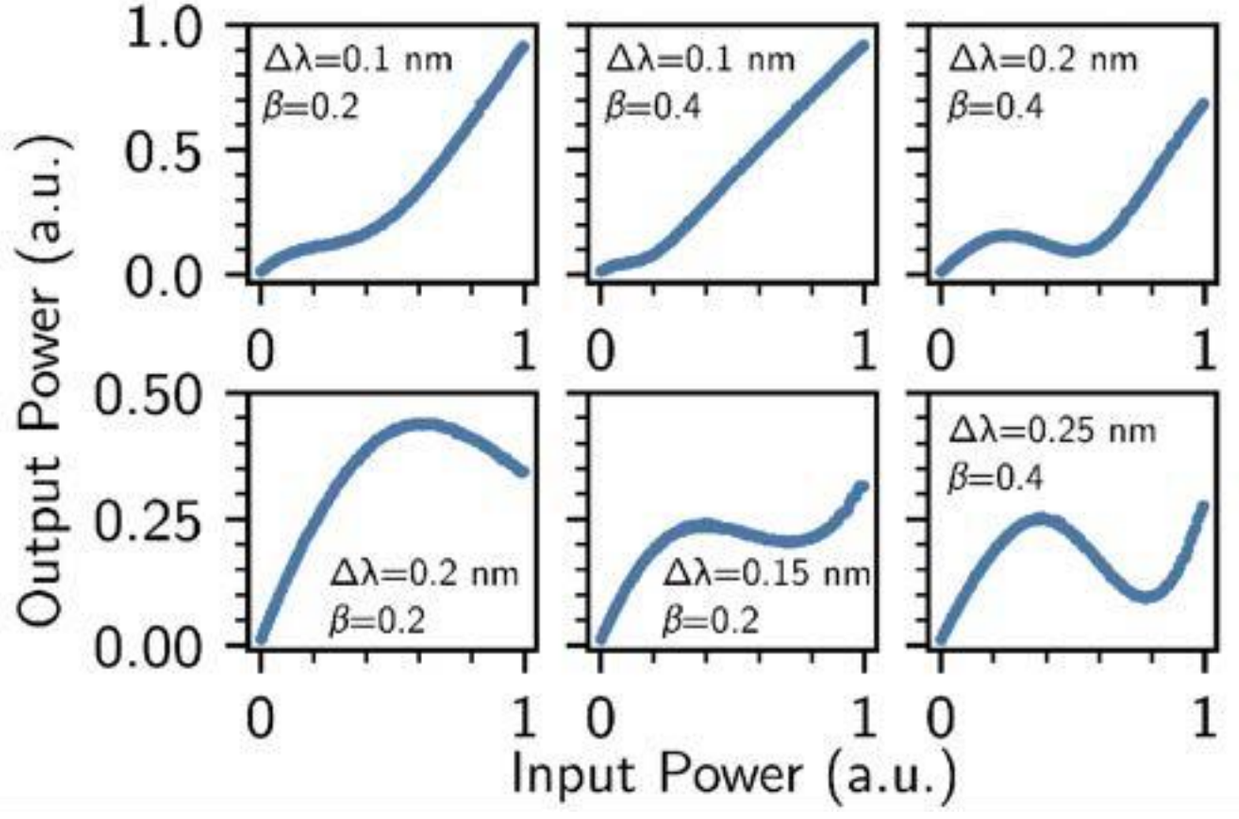
Preprint manuscript available at https://arxiv.org/abs/2208.01623

Nonlinear optical function unit





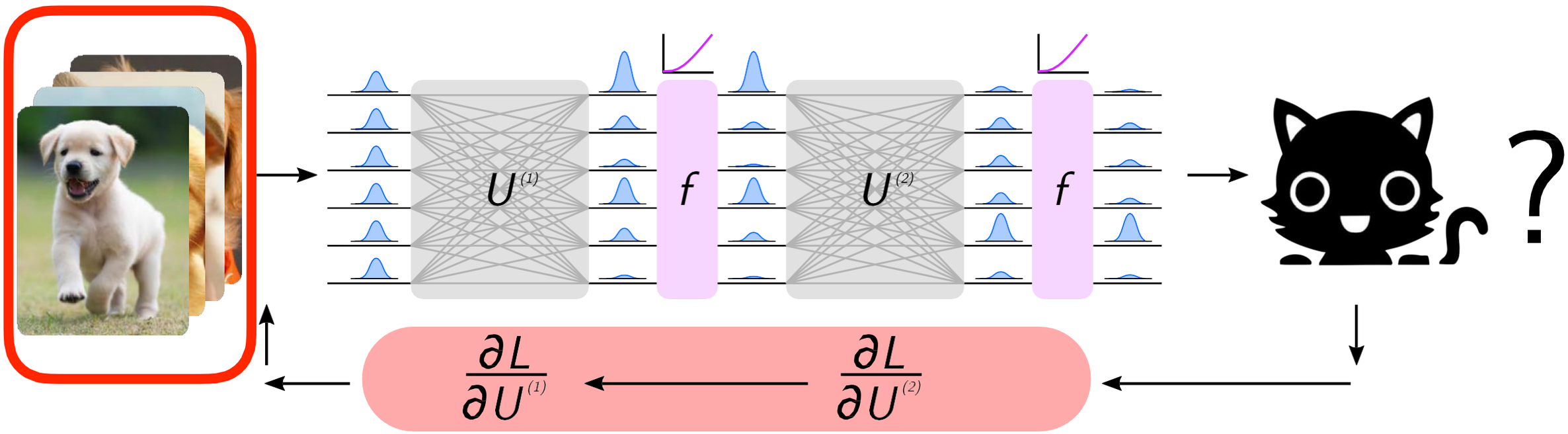




Coherent optical nonlinear functions without digitization or amplification

Preprint manuscript available at https://arxiv.org/abs/2208.01623

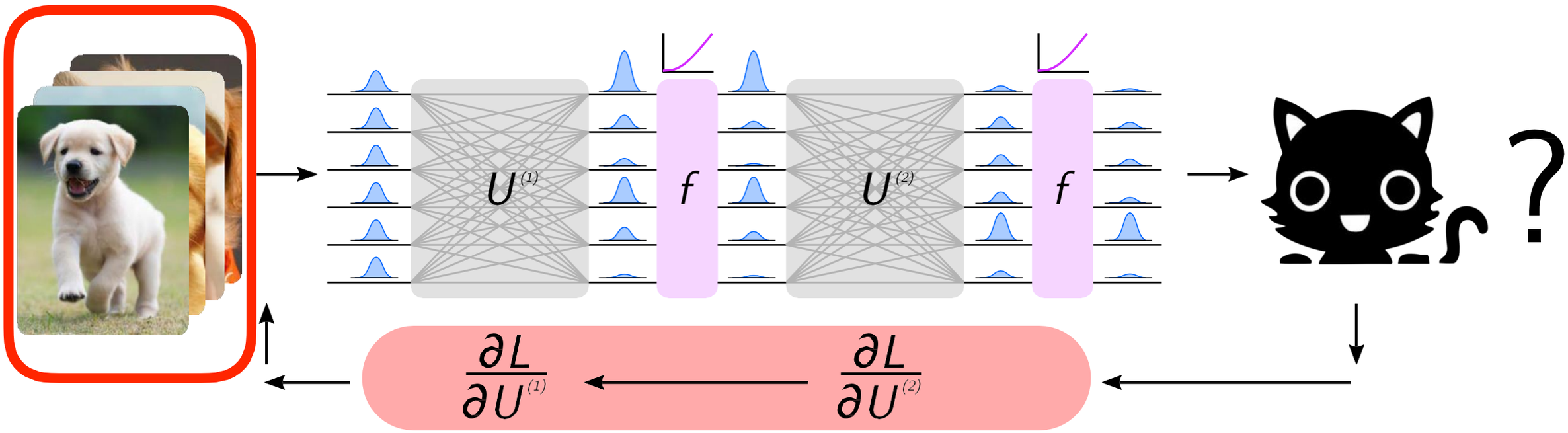
Model training benefits from optical acceleration



Model training also requires repeated forward inference.

Preprint manuscript available at https://arxiv.org/abs/2208.01623

Model training benefits from optical acceleration



Model training also requires repeated forward inference.

Optically acceleration can enable low-latency model training.

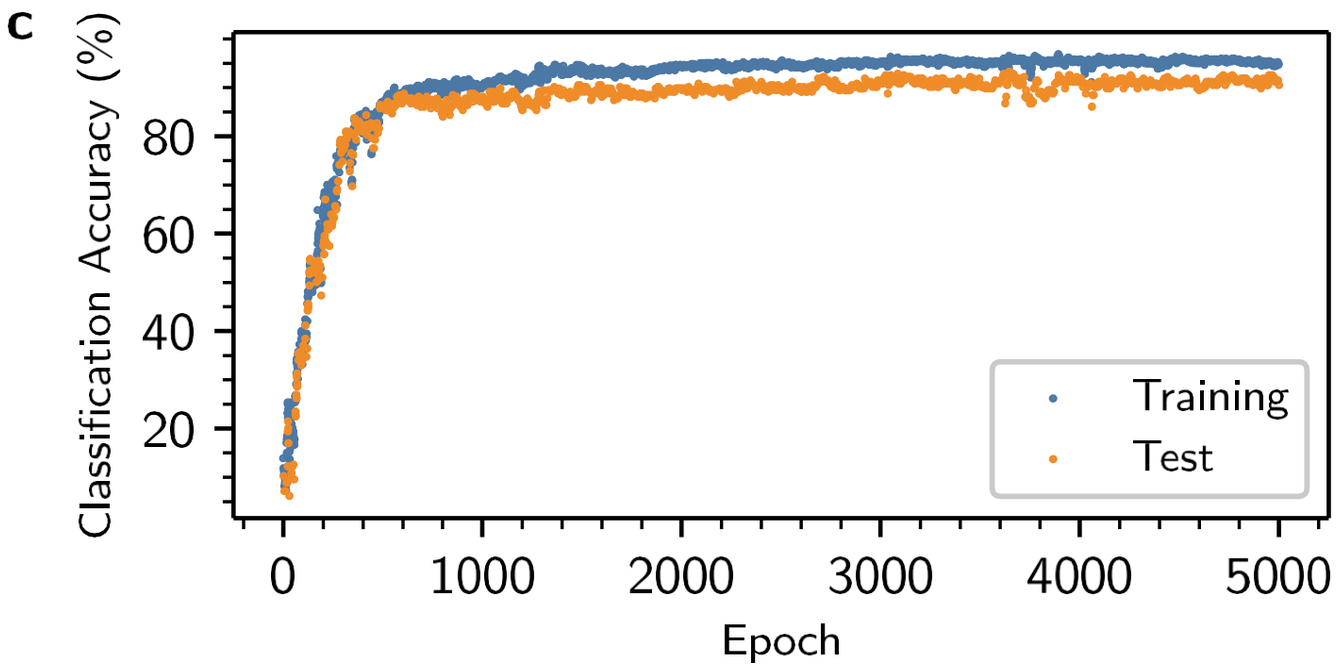
Preprint manuscript available at https://arxiv.org/abs/2208.01623

In situ training on optical hardware

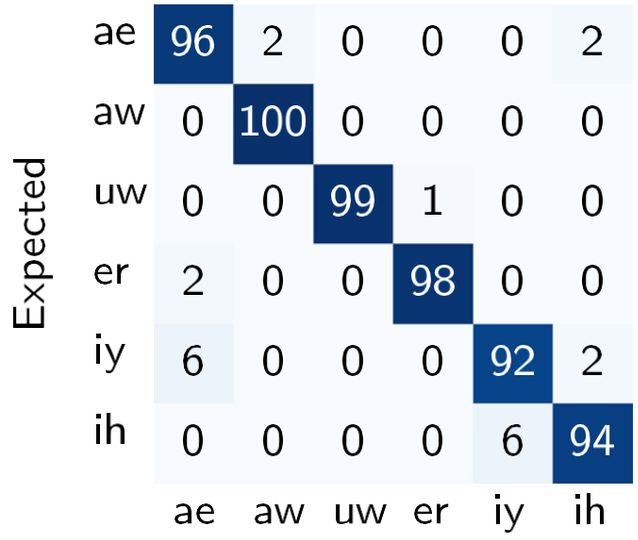
$$\boldsymbol{\Theta} + \boldsymbol{\Delta} = \boldsymbol{\Theta} + [\delta_1, \delta_2, ..., \delta_N]$$

$$\nabla_{\boldsymbol{\Delta}} \mathcal{L}(\boldsymbol{\Theta}) = \frac{\mathcal{L}(\boldsymbol{\Theta} + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\Theta} - \boldsymbol{\Delta})}{2||\boldsymbol{\Delta}||}$$

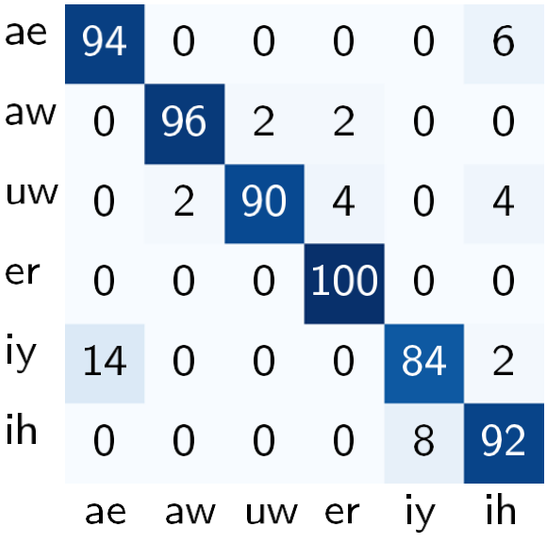
$$\boldsymbol{\Theta} \rightarrow \boldsymbol{\Theta} - \eta \nabla_{\boldsymbol{\Delta}} \mathcal{L}(\boldsymbol{\Theta}) \boldsymbol{\Delta}$$



Training Accuracy: 96.5%



Test Accuracy: 92.7%



Forward inference is **optically accelerated** during *in situ* training, which achieves the **same accuracy** as a digitally trained system.

Preprint manuscript available at https://arxiv.org/abs/2208.01623

Summary

- Single-shot optical inference in a deep neural network on a photonic chip with time-of-flight (**500 ps**) limited latency

Preprint manuscript available at https://arxiv.org/abs/2208.01623

Summary

- Single-shot optical inference in a deep neural network on a photonic chip with time-of-flight (**500 ps**) limited latency
- All-optical processing of data without digitization between layers, eliminating the latency introduced by optical-to-electronic conversion
  - Coherent matrix multiplication in the optical domain
  - Coherent optical nonlinear functions without electronic amplification

Preprint manuscript available at https://arxiv.org/abs/2208.01623

Summary

- Single-shot optical inference in a deep neural network on a photonic chip with time-of-flight (**500 ps**) limited latency
- All-optical processing of data without digitization between layers, eliminating the latency introduced by optical-to-electronic conversion
  - Coherent matrix multiplication in the optical domain
  - Coherent optical nonlinear functions without electronic amplification
- Entirely fabricated in a commercial foundry

Preprint manuscript available at https://arxiv.org/abs/2208.01623

Summary

- Single-shot optical inference in a deep neural network on a photonic chip with time-of-flight (**500 ps**) limited latency
- All-optical processing of data without digitization between layers, eliminating the latency introduced by optical-to-electronic conversion
  - Coherent matrix multiplication in the optical domain
  - Coherent optical nonlinear functions without electronic amplification
- Entirely fabricated in a commercial foundry
- *In situ* training takes advantage of near-instantaneous inference, reducing latency and power consumption of model training.

Preprint manuscript available at https://arxiv.org/abs/2208.01623

Acknowledgments

Alexander Sludds

Prof. Stefan Krastanov

Dr. Ryan Hamerly

Dr. Nicholas Harris

Dr. Darius Bunandar

Dr. Matthew Streshinsky

Dr. Michael Hochberg

Prof. Dirk Englund

Preprint manuscript available at https://arxiv.org/abs/2208.01623