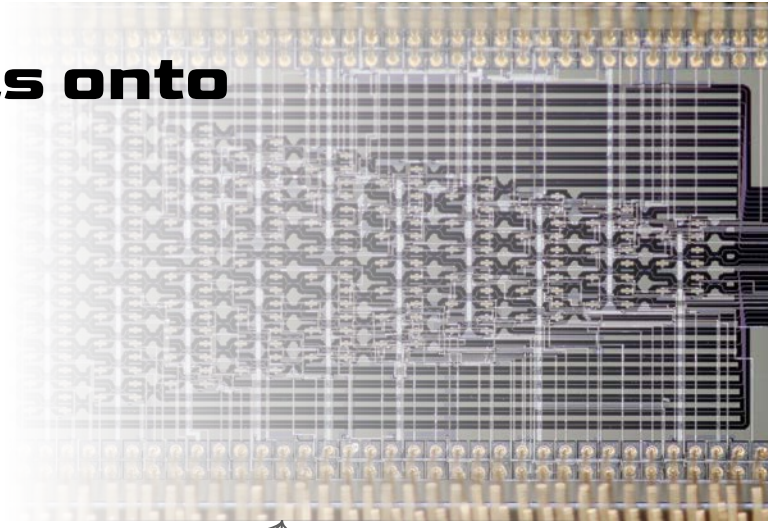
Compiling deep learning tasks onto (quantum-) optical systems



MIT AI Hardware Program Symposium | 2024-05-01

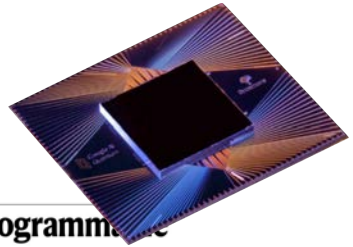
Dirk Englund ∈ {RLE, MTL, EECS}@MIT

Quantum Computing & Networking

ML signal processing

A Renaissance in Computing

1. Quantum Computing



Quantum computers use states in superposition: 000,001,..., 111

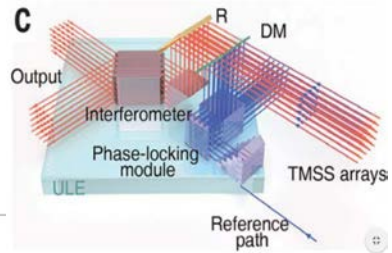
Article

Quantum supremacy using a programmable superconducting processor

https://doi.org/10.1038/s41586-019-1666-5 Frank Arute[1], Kunal Arya[1], Ryan Babbush[1], Dave Bacon[1], Joseph C. Bardin[1,2], Rami Barends[1], Rupak Biswas[3], Sergio Boixo[1], Fernando G. S. L. Brandao[1,4], David A. Buell[1], Brian Burkett[1],

Received: 22 July 2019

Google (2019)



Science

QUANTUM SIMULATION

Quantum optimization of maximum independent set using Rydberg atom arrays

S. Ebadi[1†], A. Keesling[1,2†], M. Cain[1†], T. T. Wang[1], H. Levine[1‡], D. Bluvstein[1], G. Semeghini[1], A. Omran[1,2], J.-G. Liu[1,2], R. Samajdar[1], X.-Z. Luo[1,3,4], B. Nash[5], X. Gao[1], B. Barak[5], E. Farhi[6,7], S. Sachdev[1,8], N. Gemelke[9], L. Zhou[1,9], S. Choi[1], H. Pichler[10,11], S.-T. Wang[9], M. Greiner[1*], V. Vuletić[12,*], M. D. Lukin[1,*]

Quantum computational advantage using photons

Han-Sen Zhong[1,2*], Hui Wang[1,2*], Yu-Hao Deng[1,2*], Ming-Cheng Chen[1,2*], Li-Chao Peng[1,2], Yi-Han Luo[1,2], Jian Qin[1,2], Dian Wu[1,2], Xing Ding[1,2], Yi Hu[1,2], Peng Hu[2], Xiao-Yan Yang[2], Wei-Jun Zhang[2], Hao Li[2], Yuxuan Li[1], Xiao Jiang[1,2], Lin Gan[1], Guangwen Yang[1], Lixing You[2], Zhen Wang[2], Li Li[1,2], Nai-Le Liu[1,2], Chao-Yang Lu[1,2], Jian-Wei Pan[1,2]†

Article

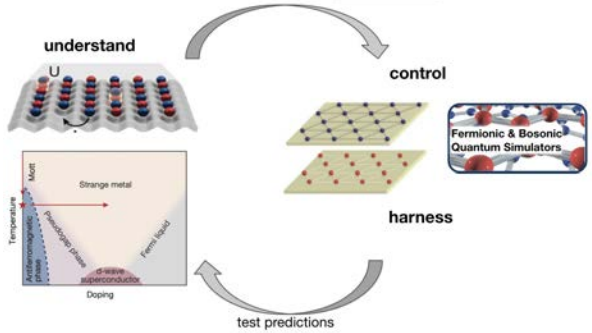
Quantum computational advantage with a programmable photonic processor

https://doi.org/10.1038/s41586-022-04725-x Lars S. Madsen[1,2], Fabian Laudenbach[3,4], Mohsen Falamarzi Askarani[1,2], Fabien Rortais[1], Trevor Vincent[1], Jacob F. F. Bulmer[1], Filippo M. Miatto[1], Leonhard Neuhaus[1], Lukas G. Helt[1], Matthew J. Collins[1], Adriana E. Lita[1], Thomas Gerrits[1], Sae Woo Nam[1], Varun D. Vaidya[1], Matteo Menotti[1], Ish Dhand[2], Zachary Vernon[2], Nicolás Quesada[2,5] & Jonathan Lavoie[2,5]

Received: 12 November 2021

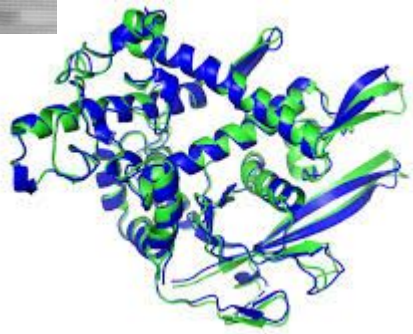
Accepted: 5 April 2022

*Noisy Intermediate-Scale Quantum (John Preskill)



2. Machine learning

Natural language processing



Protein folding [AlphaFold by Deepmind]

See also (different topic) of AI for quantum discovery - Tegmark, Soljacic, Lewenstein, Krenn, Marquardt, ..

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

Hardware

Applications

- A Renaissance in Physics of Computing ?

quantum information

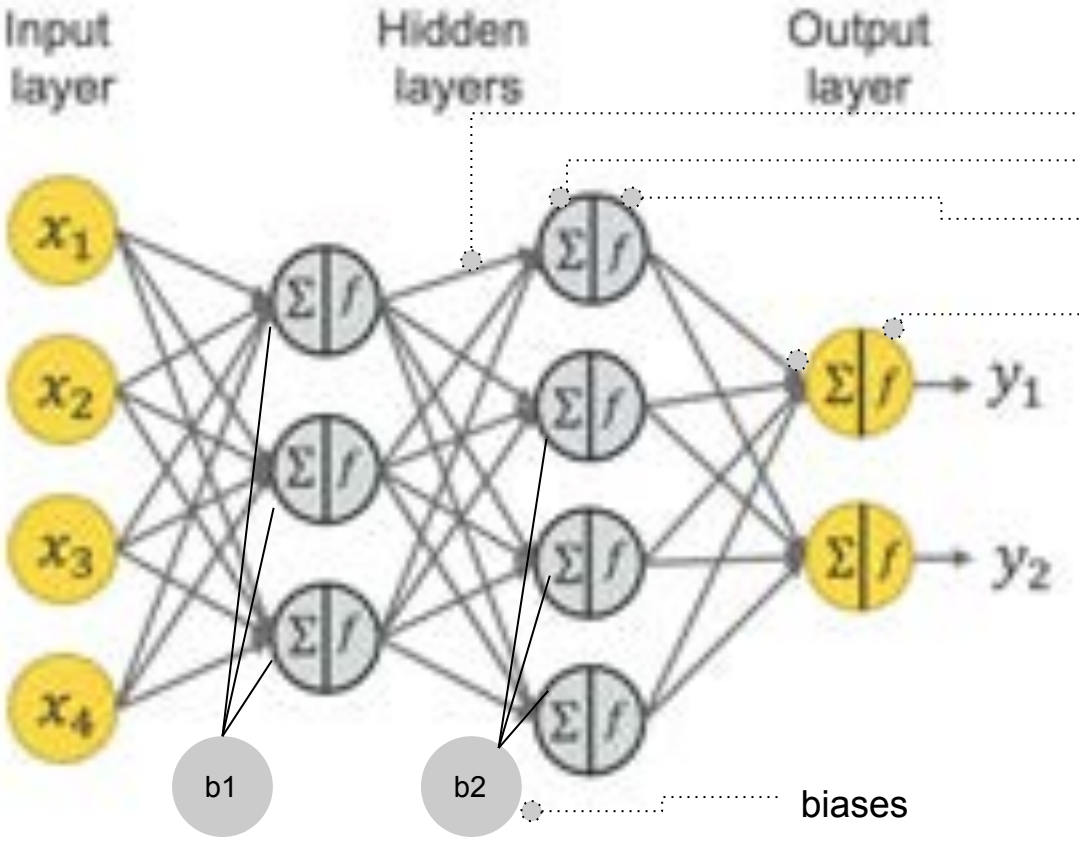
machine learning

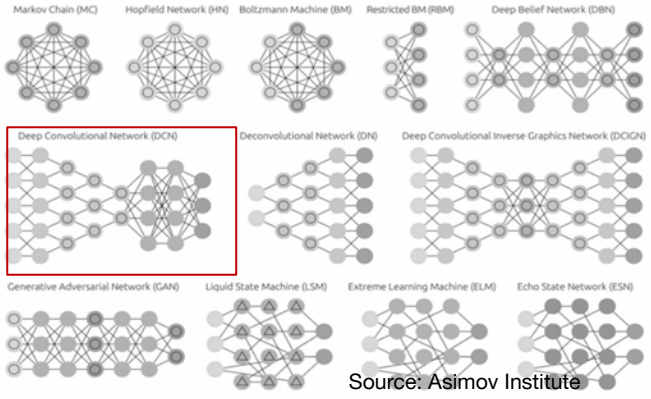
Review: G Wetzstein et al., Nature 588 (2020)

883

Deep neural network (DNN)

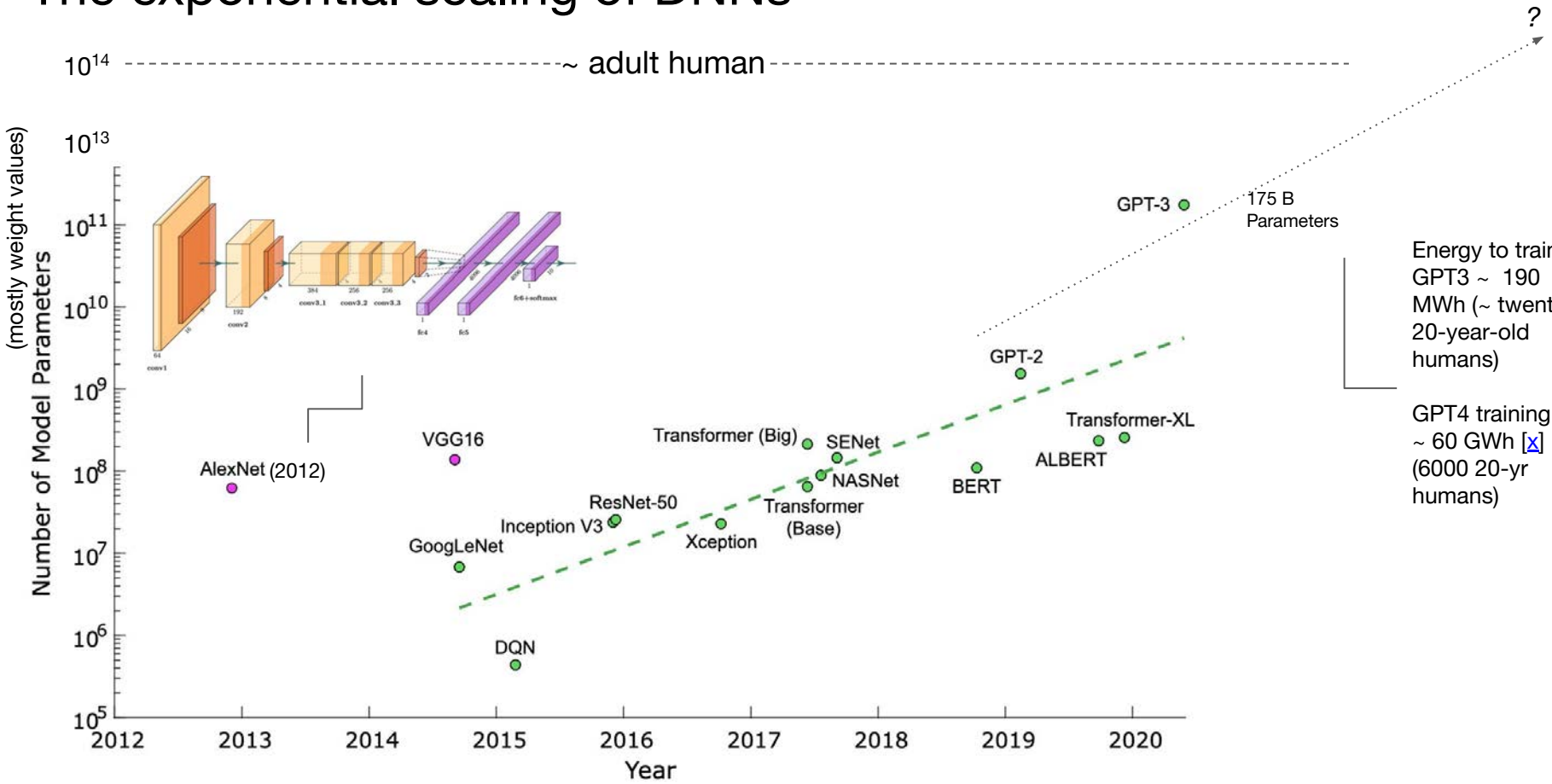
Definition: An artificial neural network with multiple layers between the input and output layers.





Source: Asimov Institute

The exponential scaling of DNNs



Energy to train GPT3 ~ 190 MWh (~ twenty 20-year-old humans)

GPT4 training ~ 60 GWh [x] (6000 20-yr humans)

Taken from L Bernstein*, A Sludds*, R Hamerly, V Sze, J Emer, D Englund, Sci Reports (2021) 11:3144

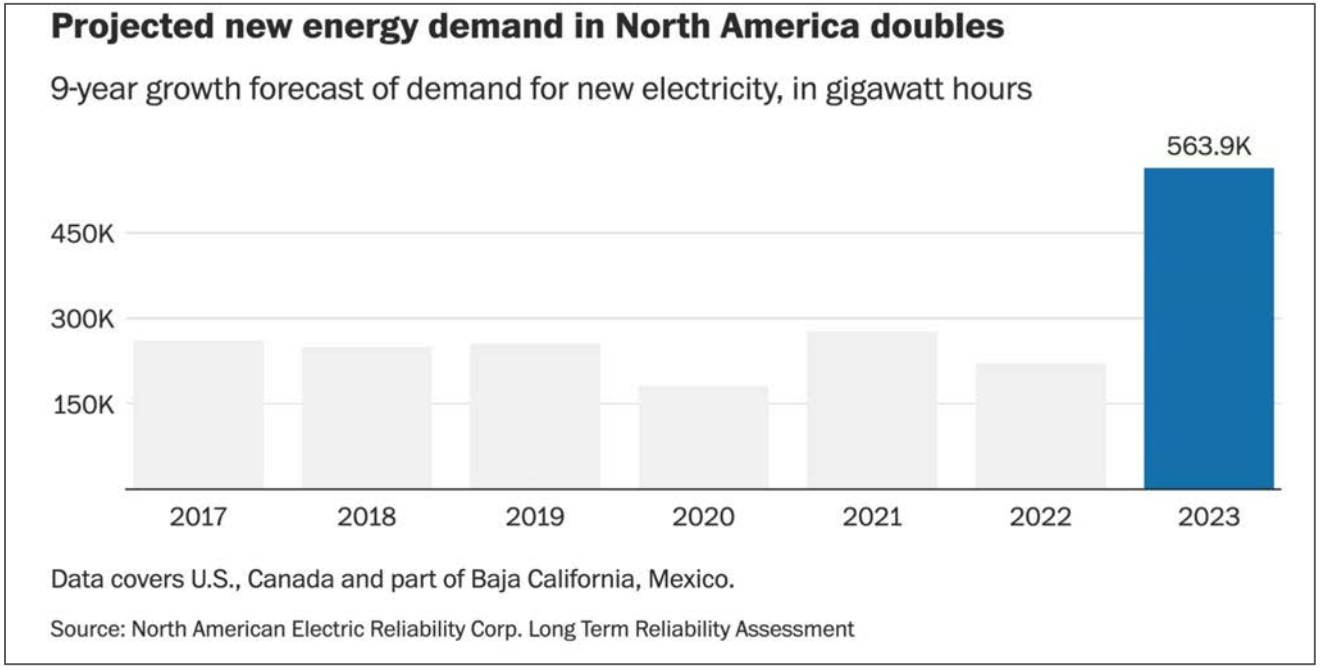
The Washington Post
Democracy Dies in Darkness

Amid explosive demand, America is running out of power

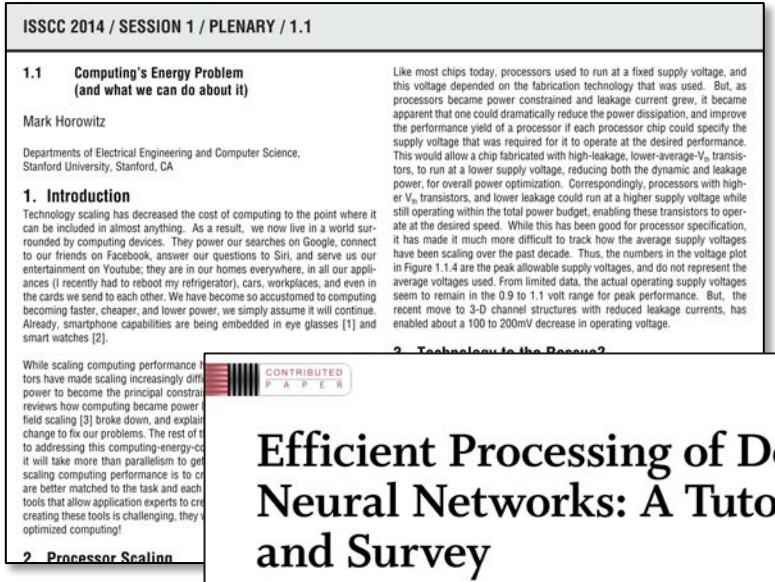
AI and the boom in clean-tech manufacturing are pushing America's power grid to the brink. Utilities can't keep up.

By Evan Halper

March 7, 2024 at 6:05 a.m. EST



Designing Systems to Tackle Power & Data Movement in DNNs



ISSCC 2014 / SESSION 1 / PLENARY / 1.1

1.1 Computing's Energy Problem (and what we can do about it)

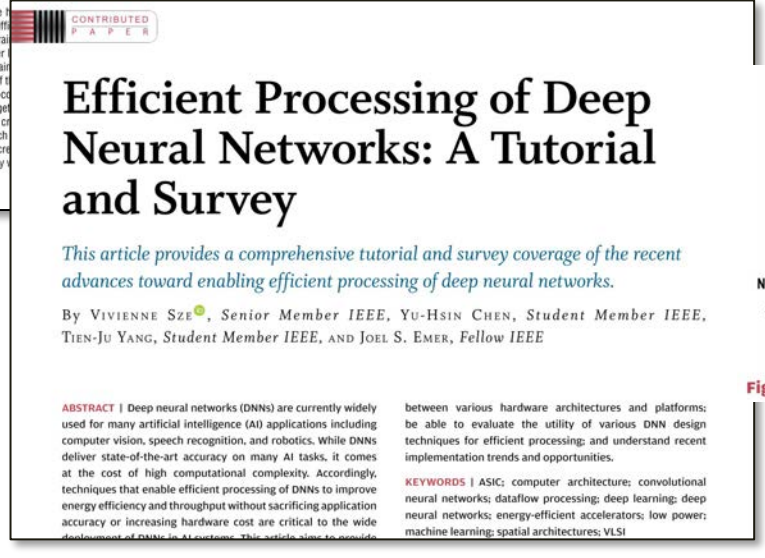
Mark Horowitz

Departments of Electrical Engineering and Computer Science, Stanford University, Stanford, CA

1. Introduction

Technology scaling has decreased the cost of computing to the point where it can be included in almost anything. As a result, we now live in a world surrounded by computing devices. They power our searches on Google, connect to our friends on Facebook, answer our questions to Siri, and serve us our entertainment on Youtube; they are in our homes everywhere, in all our appliances (I recently had to reboot my refrigerator), cars, workplaces, and even in the cards we send to each other. We have become so accustomed to computing becoming faster, cheaper, and lower power, we simply assume it will continue. Already, smartphone capabilities are being embedded in eye glasses [1] and smart watches [2].

Like most chips today, processors used to run at a fixed supply voltage, and this voltage depended on the fabrication technology that was used. But, as processors became power constrained and leakage current grew, it became apparent that one could dramatically reduce the power dissipation, and improve the performance yield of a processor if each processor chip could specify the supply voltage that was required for it to operate at the desired performance. This would allow a chip fabricated with high-leakage, lower-average-V_{th} transistors, to run at a lower supply voltage, reducing both the dynamic and leakage power, for overall power optimization. Correspondingly, processors with higher V_{th} transistors, and lower leakage could run at a higher supply voltage while still operating within the total power budget, enabling these transistors to operate at the desired speed. While this has been good for processor specification, it has made it much more difficult to track how the average supply voltages have been scaling over the past decade. Thus, the numbers in the voltage plot in Figure 1.1.4 are the peak allowable supply voltages, and do not represent the average voltages used. From limited data, the actual operating supply voltages seem to remain in the 0.9 to 1.1 volt range for peak performance. But, the recent move to 3-D channel structures with reduced leakage currents, has enabled about a 100 to 200mV decrease in operating voltage.



CONTRIBUTED PAPER

Efficient Processing of Deep Neural Networks: A Tutorial and Survey

This article provides a comprehensive tutorial and survey coverage of the recent advances toward enabling efficient processing of deep neural networks.

By Vivienne Sze, *Senior Member IEEE*, Yu-Hsin Chen, *Student Member IEEE*, Tien-Ju Yang, *Student Member IEEE*, and Joel S. Emer, *Fellow IEEE*

ABSTRACT | Deep neural networks (DNNs) are currently widely used for many artificial intelligence (AI) applications including computer vision, speech recognition, and robotics. While DNNs deliver state-of-the-art accuracy on many AI tasks, it comes at the cost of high computational complexity. Accordingly, techniques that enable efficient processing of DNNs to improve energy efficiency and throughput without sacrificing application accuracy or increasing hardware cost are critical to the wide

between various hardware architectures and platforms; be able to evaluate the utility of various DNN design techniques for efficient processing; and understand recent implementation trends and opportunities.

KEYWORDS | ASIC; computer architecture; convolutional neural networks; dataflow processing; deep learning; deep neural networks; energy-efficient accelerators; low power; machine learning; spatial architectures; VLSI

Takes more energy to fetch longer distances in memory

- Off-chip >> on-chip
- Large RAM > Small RAM ($E_{bit} \propto L \propto (N_{bit})^{0.5}$)
- Goal: max. use of small register memories → "local data reuse"

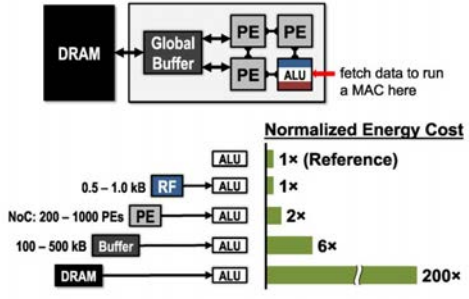


Fig. 22. *Memory hierarchy and data movement energy [82].*

V. Sze, Proc. IEEE 105(12), 2295 (2017)

Integer	
Add	
8 bit	0.03pJ
32 bit	0.1pJ
Mult	
8 bit	0.2pJ
32 bit	3.1pJ

FP	
FAdd	
16 bit	0.4pJ
32 bit	0.9pJ
FMult	
16 bit	1.1pJ
32 bit	3.7pJ

Memory	
Cache	(64bit)
8KB	10pJ
32KB	20pJ
1MB	100pJ
DRAM	1.3-2.6nJ

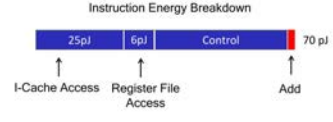
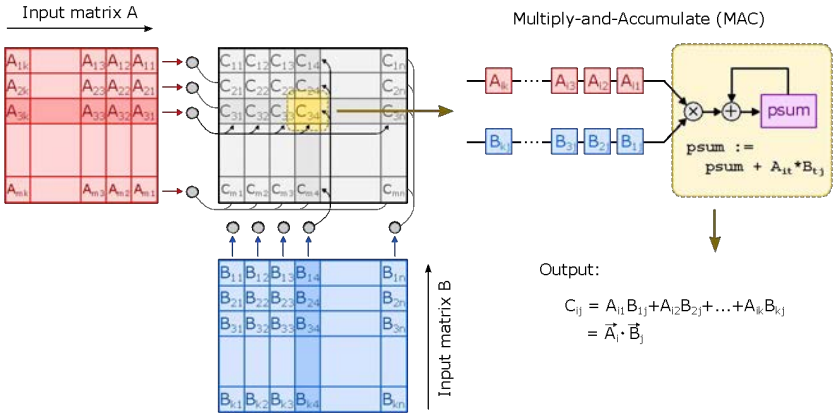


Figure 1.1.9: Rough energy costs for various operations in 45nm 0.9V.

M. Horowitz, ISSCC 2014, pp.10-14

slide: Ryan Hamerly

Systolic Array architecture is optimized for general matrix multiplier



Two key limiters to the TPU:

- Power → Heat dissipation
- Interconnects / Data Movement

slide credit: R Hamerly

In-Datacenter Performance Analysis of a Tensor Processing Unit

Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa,

.. Google team: >cs>arXiv:1704.04760

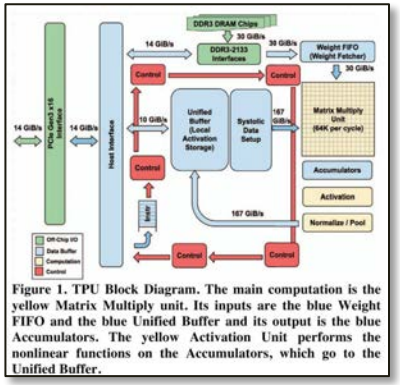


Figure 1. TPU Block Diagram. The main computation is the yellow Matrix Multiply unit. Its inputs are the blue Weight FIFO and the blue Unified Buffer and its output is the blue Accumulators. The yellow Activation Unit performs the nonlinear functions on the Accumulators, which go to the Unified Buffer.

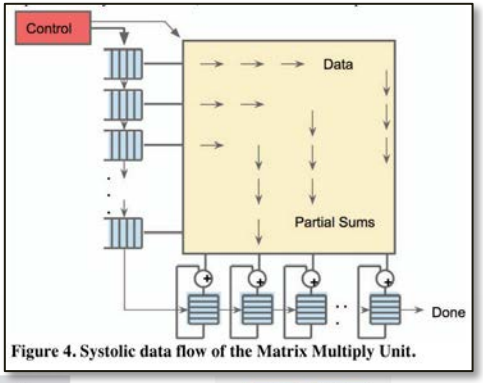
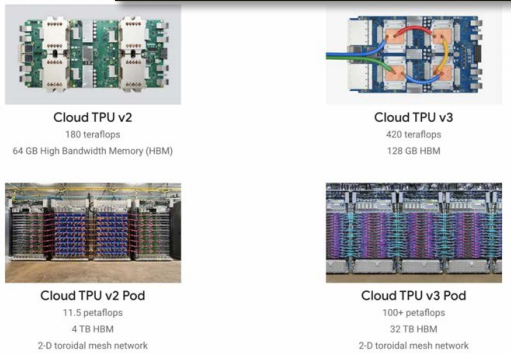
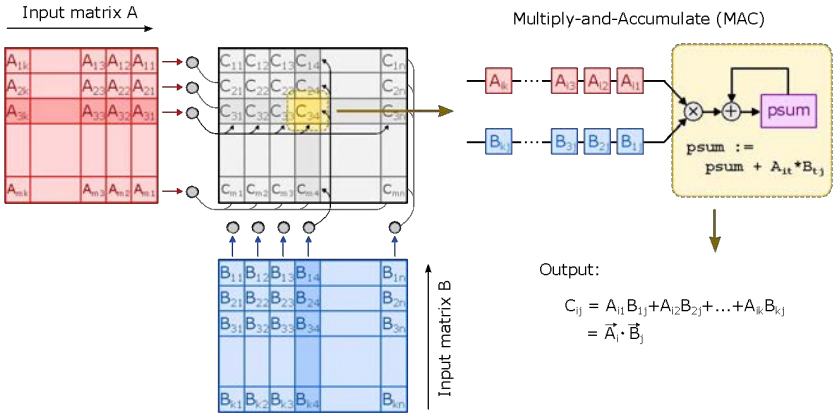


Figure 4. Systolic data flow of the Matrix Multiply Unit.

Google TPU cloud offerings:



Systolic Array architecture is optimized for general matrix multiplier



In-Datacenter Performance Analysis of a Tensor Processing Unit

.. Google team: >cs>arXiv:1704.04760

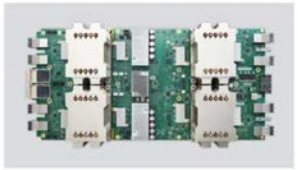
Google TPU cloud offerings:

Two key limiters to the TPU:

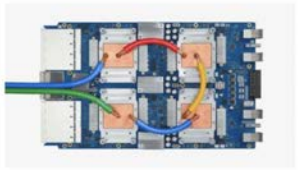
- Power ⟶ Heat dissipation
- Interconnects / Data Movement

Despite all the optimizations, chips are dominated by data movement costs, not raw processing!

Cloud TPU offering

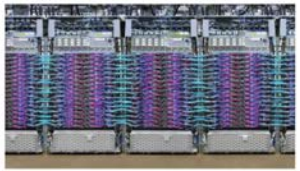


Cloud TPU v2
180 teraflops
64 GB High Bandwidth Memory (HBM)



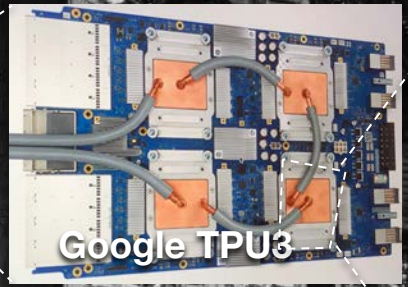
Cloud TPU v3
420 teraflops
128 GB HBM

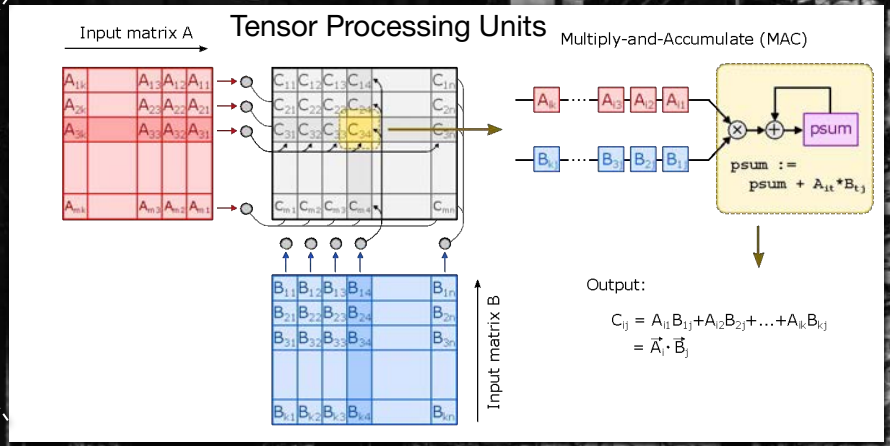
Cloud TPU v2 Pod
11.5 petaflops
4 TB HBM
2-D toroidal mesh network



Cloud TPU v3 Pod
100+ petaflops
32 TB HBM
2-D toroidal mesh network

slide credit: R Hamerly

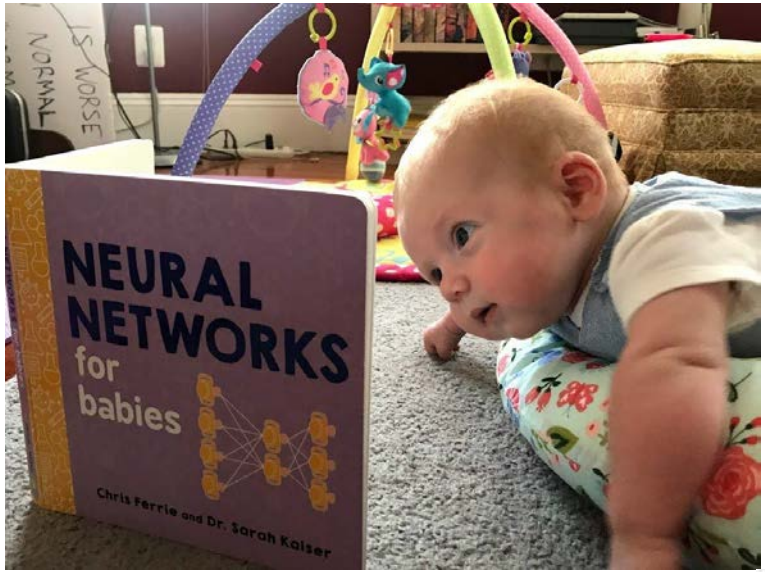




Two key limiters: Power still dominated by data movement costs, not transistors

* Google team: >cs>arXiv:1704.04760

Information processing architectures

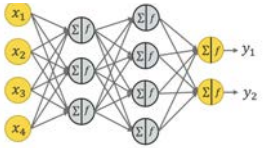


Non-boolean logic

Boolean

Google Sheets ✅

?



?

Acknowledgements (ML portion)

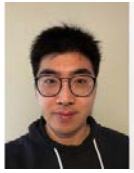
Positions available in theory and experiment: http://www.rle.MIT.edu/qp

MIT Quantum Photonics Group: Dr. Ryan Hamerly (NTT, MIT)

PhD: Saumil Bandyopadhyay, Nick Harris (--> LightMatter), Darius Bunandar (--> LightMatter), Mihika Prabhu, Chris Panuski, Liane Bernstein, Alex Sludds, Ronald Davis

Postdocs: Stefan Krastanov, Sri Krishna Vadlamani, Zaijun Chen (--> USC)


Saumil Bandyopadhyay


Prof. Zaijun Chen


Dr. Nicholas Harris


Dr. Zhizhen Zhong


Liane Bernstein


Dr. Stefan Krastanov


Dr. Carlos Herranz


Sivan Trajtenberg-Mills


Hugo Larocque


CPT Dr Chris Panuski (--> U.S. Navy)

Collaborators on this work:

MIT: Manya Ghobadi, Joel Emer, Vivienne Sze

MITRE Corp: Gerry Gilbert, Mark Dong, Gen Clark
ARL: Dr Dashiel Vittulo, Dr Kurt Jacobs

Sandia NL: M Eichenfield
AFRL: Dr Mike Fanto, et al


Alexander Sludds


Ronald Davis III

Dr Sri Vadlamani


Prof. Manya Ghobadi


Dr. Ryan Hamerly

Funding

AFOSR AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
NSF
NTT
ANALOG DEVICES
SCIENCE TO SHAPE THE FUTURE OF THE ARMY · ARMY RESEARCH OFFICE · 1951
MITRE
HRI Honda Research Institute US
tsmc
DARPA
MTL MICROSYSTEMS TECHNOLOGY LABORATORIES

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

quantum information

machine learning

Review: G Wetzstein et al, Nature 588 (2020)

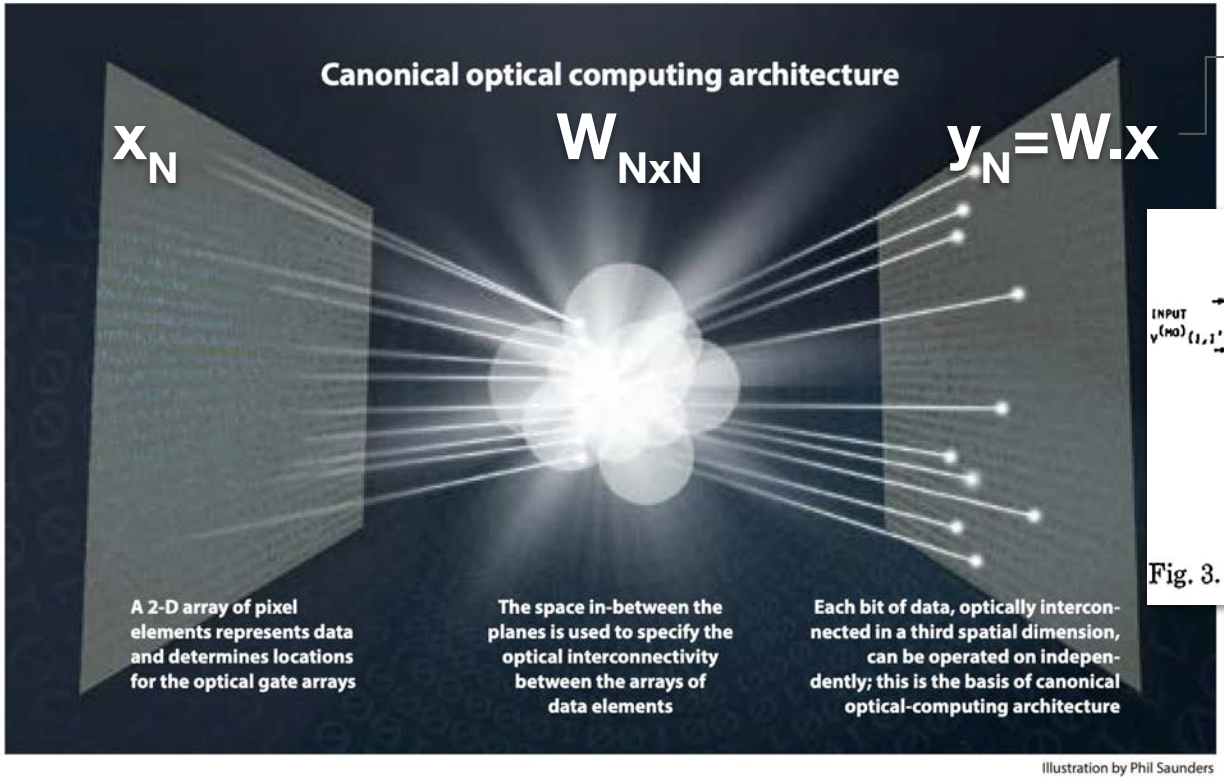
Review: W Bogaerts et al, Nature 586, 207–216

Hardware

- In-physics processing

Applications

Ex.1: “Weight-stationary” compute core



Effort scales as O(N) in optics

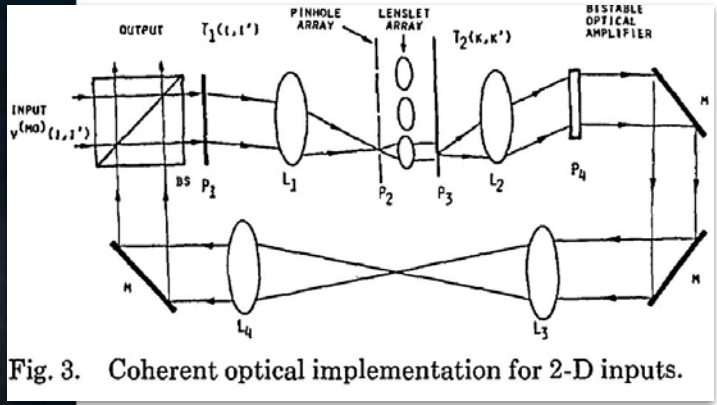


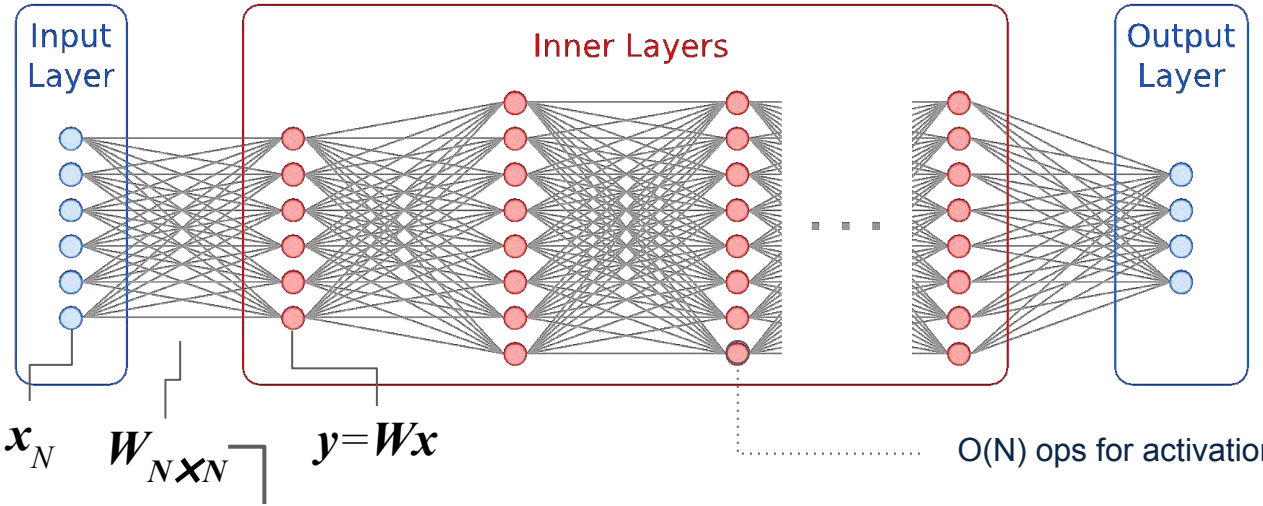
Fig. 3. Coherent optical implementation for 2-D inputs.

Demetri Psaltis and Nabil Farhat, "Optical information processing based on an associative-memory model of neural nets with thresholding and feedback," Opt. Lett. 10, 98-100 (1985)

Ravi Athale and Demetri Psaltis, OPN **27** (2016)

Ex.1:

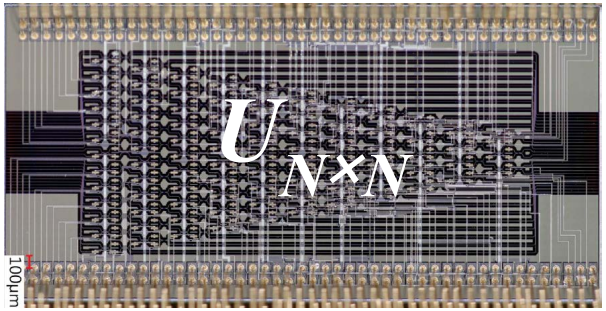
"Weight-stationary architecture" output **y**=(stationary **W**).(flying **x**)



→ MIT spinouts based on new architectures

LIGHTMATTER

LIGHTELLIGENCE



$U_{N\times N}\,x_N$ vector-matrix multiplication:

Binary computer: O(N^2) ops + memory access

Opportunity for PIC: O(N) ops, no memory access, equivalent compute of N^2
$\Delta v \sim 100^2$ (10THz)~ 10^{17} ops/sec

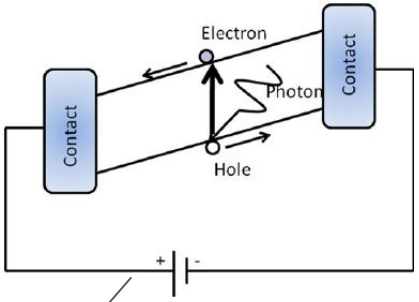
Pioneering work by Psaltis et al

Recent also by Soljacic, Gigan, Stefan Rotter, Brunner, Pan, Wetzstein, Ozcan, Fan, Marin Soljačić, P Prucnel, Cornelia Denz, D. A. B. Miller, Liboiron-Ladouceur, Pappas ..

Y Shen*, N C. Harris*, S Skirlo, M Prabhu, T Baehr-Jones, M Hochberg, X Sun, S Zhao, H Larochelle, D Englund, and M Soljacić, *Nature Photonics* 11 (2017). *equal authors

"Output-stationary architecture": spatially encoded output **y**=(flying input **X**).(flying weight matrix **W**)

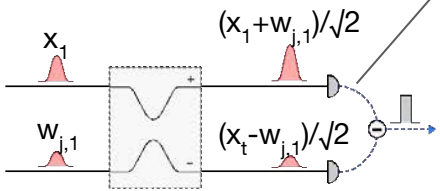
"Photoelectric multiplication"



The (quantum) process of photodetection has extreme photon-electron nonlinearity!

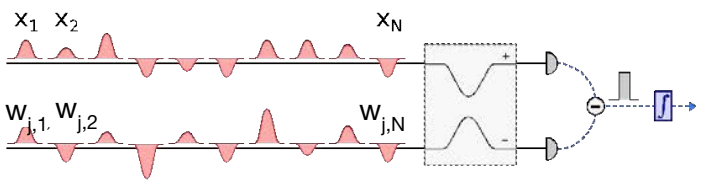
$$\langle \psi \mid E^{(-)}(\mathbf{r},t) \cdot E^{(+)}(\mathbf{r},t) \mid \psi \rangle$$

Optical Multiplication



Top PD: $I_+ = (x_t+w_{i,1})^2/2$
Bottom PD: $I_- = (x_t-w_{i,1})^2/2$
Difference: $(I_+-I_-) = 2\, x_1\, w_{j,1}$

Multiply and accumulate. Dot product.

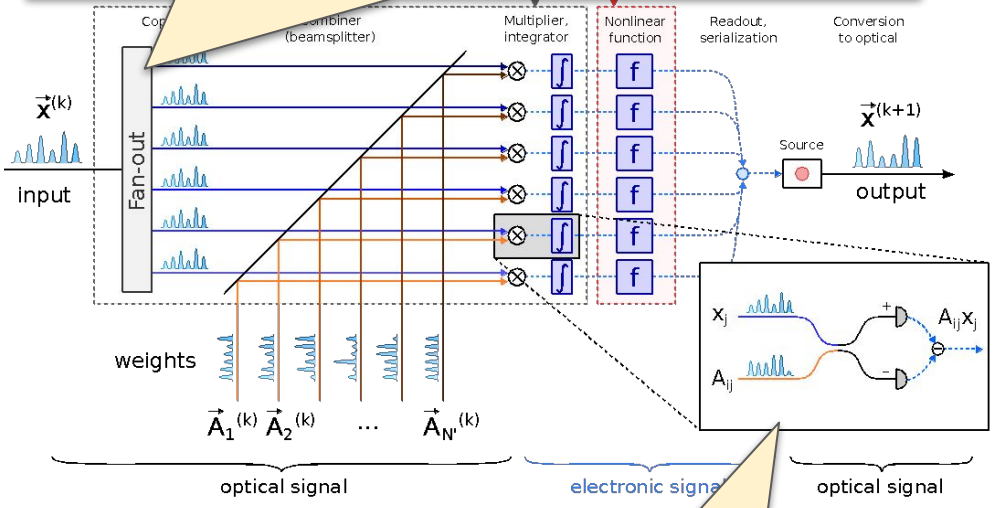


Output $\mathbf{y}_j=\Sigma_k\, x_k\, w_{j,k}=\mathrm{dot}(\mathbf{x},\mathbf{w}_j)$

R Hamerly, A Sludds, L Bernstein, M Soljacic, DE, PRX **9** (2019)

Output-stationary ONN by ***photoelectric multiplication***

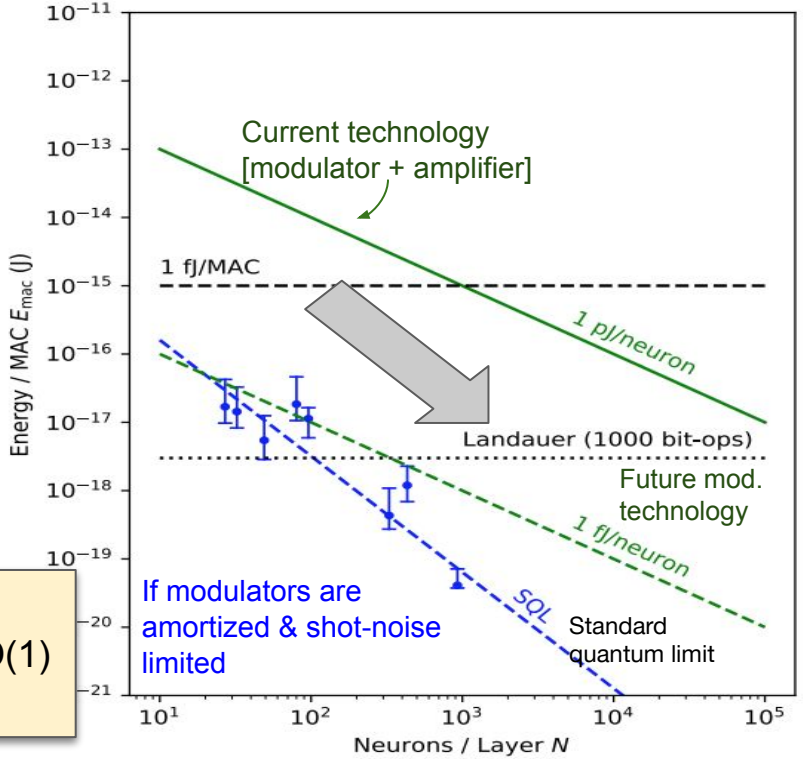
Key point #1: **Use "free" data fan-out** if possible



$$\boldsymbol{W}_{M\times N}\,\boldsymbol{x}_N \rightarrow \boldsymbol{y}_M$$

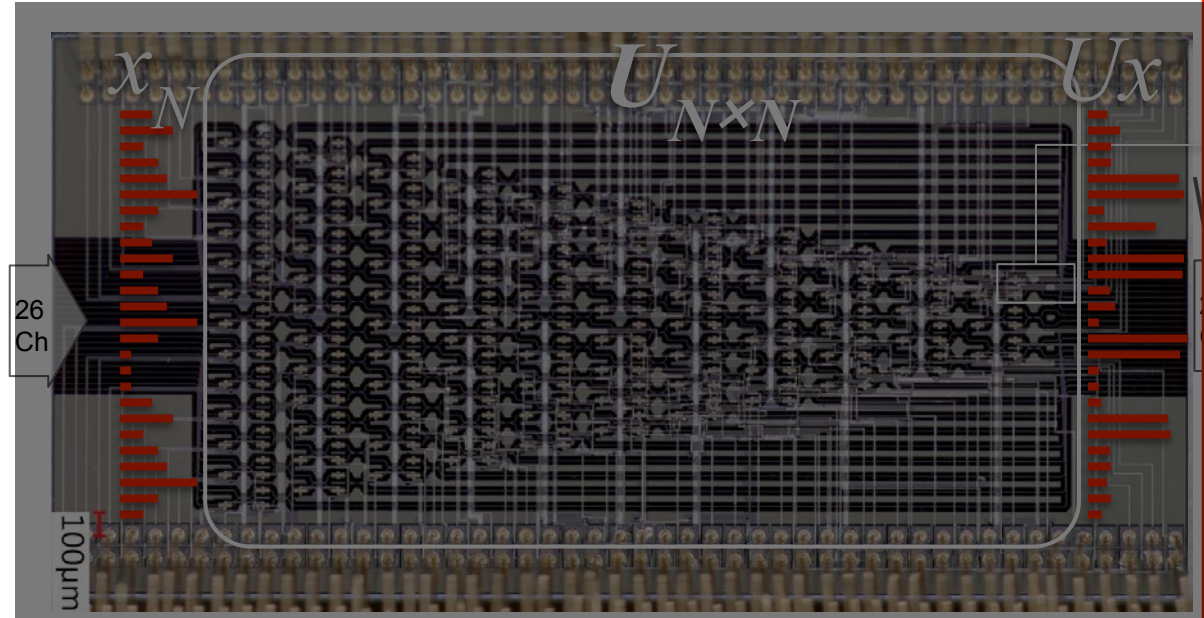
Key point #2: **In-physics computing** for dot(x,y) with O(1) resource vs. O(N) for boolean

Physics predicts von Neumann architecture would need to violate thermodynamics to keep up



R Hamerly, A Sludds, L Bernstein, M Soljacic, DE, PRX **9** (2019)

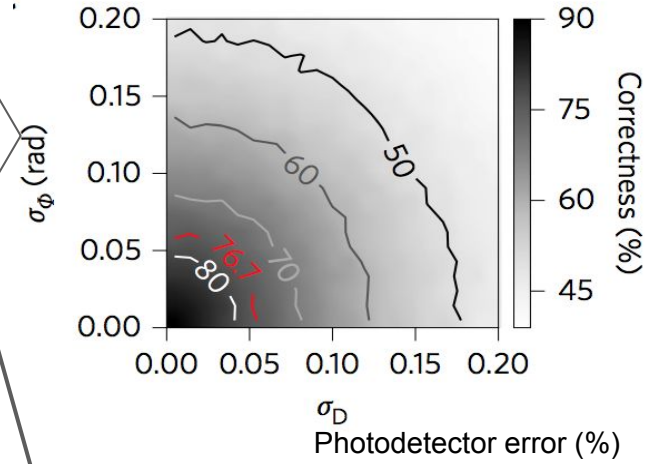
Matrix algebra in programmable photonic integrated circuits (PICs)



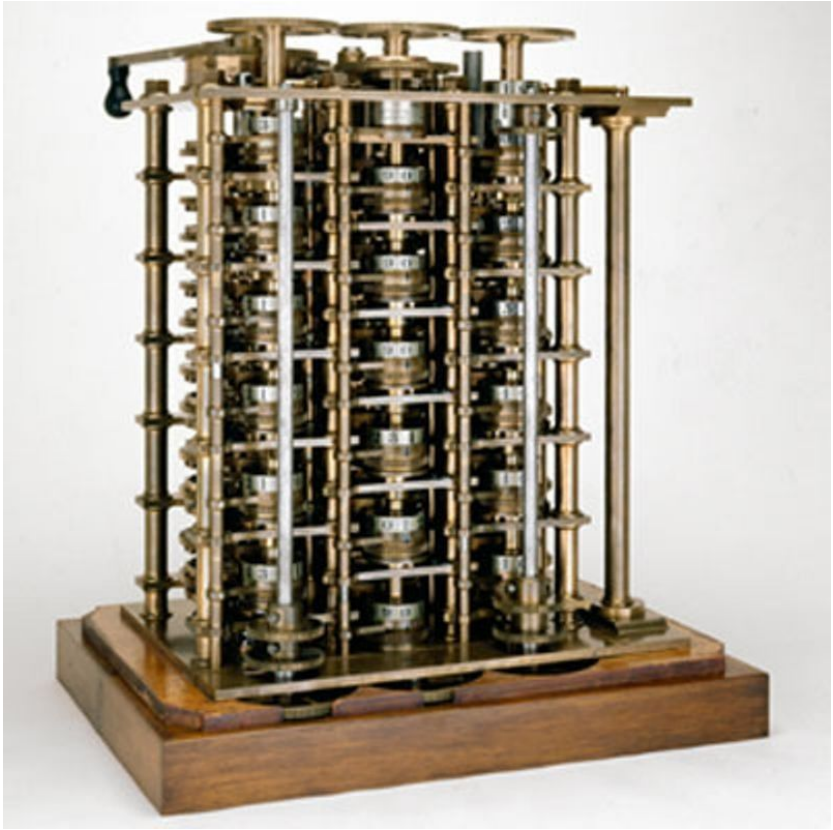
Y Shen*, N C. Harris*, et al [Englund - Soljacic groups],Nature Photonics 11 (2017).* equal authors

Error rate in vowel recognition was limited by fidelity of matrix-vector transformation → can we fix this?

Gaussian noise distribution of phase shifters, in radians



Scalability of analog systems?



C. Babbage “difference engine” (1832)

$\mathbf{y}_j=\Sigma_k x_k w_{j,k}=\mathrm{dot}(\mathbf{x},\mathbf{w}_j)$

A CONTINUOUS INTEGRAPH.*

BY

V. BUSH, F. D. GAGE, and H. R. STEWART.
Electrical Engineering Department, Massachusetts Institute of Technology.

ABSTRACT.

A MECHANICAL integraph has been developed which plots continuously the integral of the product of two functions. It uses the principle of the electrical integrating watthour-meter combined with a moving table. Errors of the machine have been reduced to an average of 1 per cent. for common uses. By cross-connecting the device in a simple mechanical way, it is possible to solve cer-

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing
- EC in hardware compute, training: → “error-free optics”
- EC in ML in-situ training?

Review: G Wetzstein et al, Nature 588 (2020)

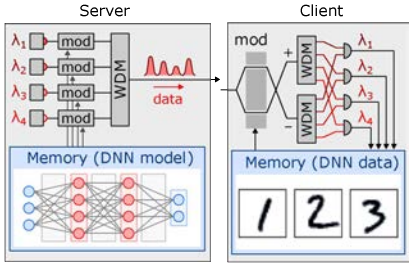
Review: W Bogaerts et al, Nature 586, 207–216 (2020)

S. Bandyopadhyay, Optica 8, 1247 (2021)

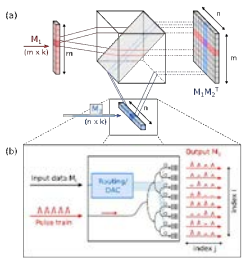
R Hamerly et al, Nature Comm. 13 (2023)

Applications

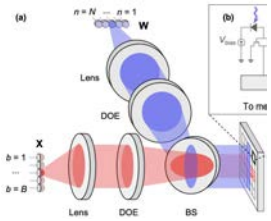
The impact of hardware errors



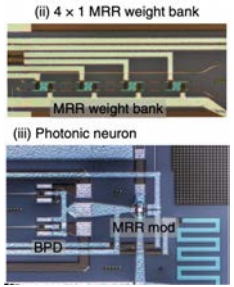
A. Sludds, OFC 2022



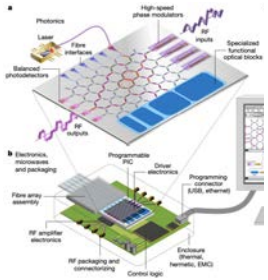
RH, PRX 2019



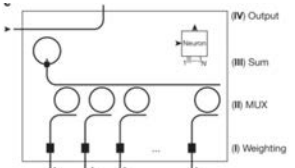
L. Bernstein, Sci. Rep. 2021



Huang, Nat. Elec. 2021



Bogaerts, Nature 2020



Feldman et al, *Nature* **569** (2019)

"Output-stationary" shallow architectures with time-frequency encoded weights

Light passes through O(1) components
No cascading of hardware errors
Error-cancellation by calibration

Weight-stationary "deep" circuits

Light passes through O(N) components
Hardware errors can cascade -- is that a show-stopper?

Error correction techniques in Mach Zehnder Meshes

Numerical Optimization

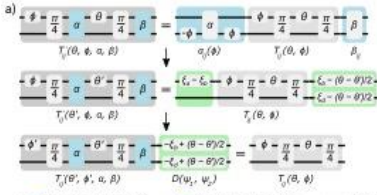
Nonlinear optimizer: eg minimize norm $|U - U_0|$

Burgwal et al., OE 25, 28236 (2017)
Mower et al., PRA 92, 032322 (2015)
Pai et al., PRApp 11, 064044 (2019)

(+) Works for arbitrary meshes
(–) Slow
(–) Needs pre-calibration (knowledge of MZI errors)

"Local" EC

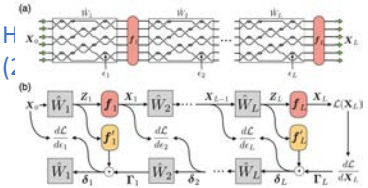
Given errors, find (θ,φ) that fix them per-MZI.



(+) Arbitrary meshes
(+) Fast
(–) Needs pre-calibration

In-situ Training

Adjoint-related method. Send input & weight gradient, get gradients in internal detectors.

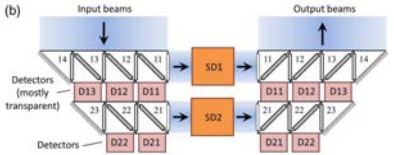


(+) Arbitrary meshes
(–) Slow, per-device training
(+) No pre-calibration
(–) Additional $O(N^2)$ detectors

Progressive Methods

Adjust MZIs one-by-one to match certain I/O modes.

Miller et al., OE 21, 6360 (2013)
Miller et al., Phot Res 1, 1 (2013)



(–) Only triangular meshes
(+) Fast
(+) No pre-calibration
(–) Additional $O(N^2)$ detectors

Local Error Correction

S. Bandyopadhyay, *Optica* 8, 1247 (2021)

Research Article

OPTICA

Hardware error correction for programmable photonics

Saumil Bandyopadhyay,[1,*] Ryan Hamerly,[1,2] and Dirk Englund[1]

[1]Research Laboratory of Electronics, MIT, 50 Vassar Street, Cambridge, Massachusetts 02139, USA
[2]NTT Research Inc., PHI Laboratories, 940 Stewart Drive, Sunnyvale, California 94085, USA
*Corresponding author: saumilb@mit.edu

Received 2 March 2021; revised 1 July 2021; accepted 18 August 2021 (Doc. ID 424052); published 27 September 2021

Programmable photonic circuits of reconfigurable interferometers can be used to implement arbitrary operations on optical modes, providing a flexible platform for accelerating tasks in quantum simulation, signal processing, and artificial intelligence. A major obstacle to scaling up these systems is static fabrication error, where small component errors within each device accrue to produce significant errors within the circuit computation. Mitigating this error usually requires numerical optimization dependent on real-time feedback from the circuit, which can greatly limit the scalability of the hardware. Here we present a deterministic approach to correcting circuit errors by locally correcting hardware errors within individual optical gates. We apply our approach to simulations of large scale optical neural networks and infinite impulse response filters implemented in programmable photonics, finding that they remain resilient to component error well beyond modern day process tolerances. Our results highlight a potential way to scale up programmable photonics to hundreds of modes with current fabrication processes.

https://doi.org/10.1364/OPTICA.424052

1. INTRODUCTION

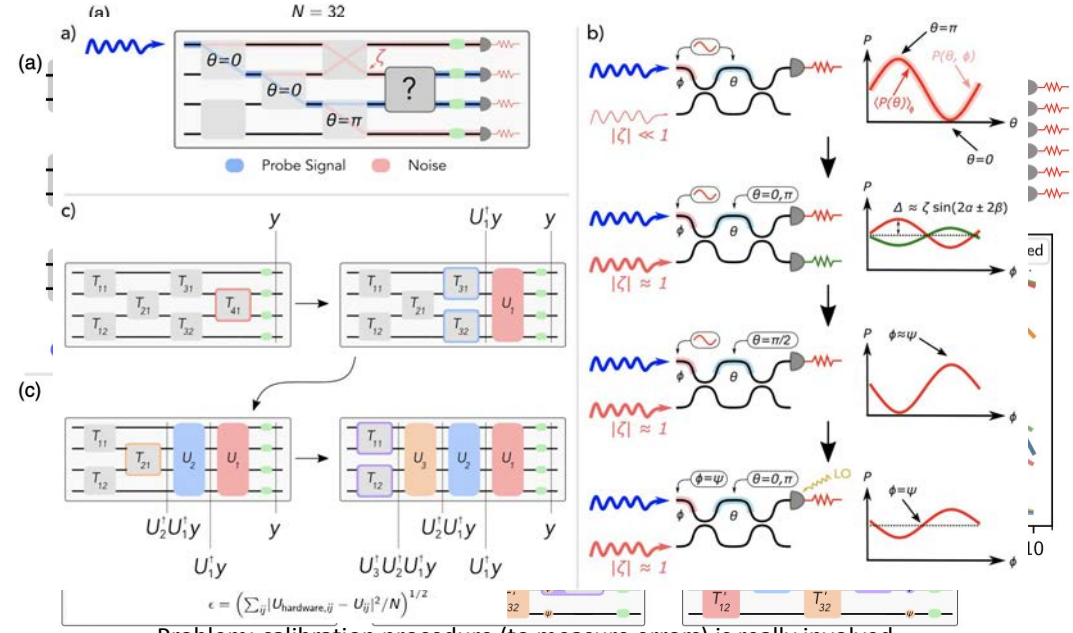
Integrated photonics is a key technology for optical communications and is advancing rapidly for applications in sensing, metrology, signal processing, and computation. Programmable photonic circuits of optical interferometers, which can implement arbitrary filters and passively compute matrix operations on optical modes, are the optical analog to field programmable gate arrays (FPGAs) and enable photonic circuits to be flexibly reconfigured post-fabrication by software [1,2]. Experimental demonstrations of these circuits have already shown working systems operating on up to tens of optical modes, which have been used to accelerate tasks in quantum simulation [3–7], mode unscrambling [8–12], signal processing [13,14], combinatorial optimization [15], and artificial intelligence [16].

While scaling up these systems to hundreds or thousands of modes would be immensely beneficial, doing so will require precise fabrication of tens of thousands of optical interferometers. Unfortunately, static component errors induced by process variation introduce errors that rapidly accrue for larger systems, limiting their usefulness for many applications. This is because the decomposition [17,18] and optimization techniques used to program these circuits assume that all of the components are ideal; thus, any component errors result in a programming of the wrong operation. Component imprecision therefore has serious implications for the future of these systems: for example, beam splitter variation as small as 2%, which is a typical wafer-level variance [19], has been shown to degrade accuracy by nearly 50% for feedforward circuits used to implement classifiers for the MNIST image recognition task [20]. Alternative programmable architectures, such as recirculating waveguide meshes consisting of triangular or hexagonal Mach–Zehnder interferometer (MZI) lattices [13,21,22], are similarly susceptible to component-induced error; device variation within these circuits introduces errors that will alter the response of phase-sensitive filters [23]. These systems' degrees of sensitivity to component variation make their control challenging when scaling up to large numbers of modes.

Hardware errors are usually compensated for with numerical optimization. A number of global optimization approaches have been proposed in the past, including nonlinear optimization [24–28], gradient descent [29], and *in situ* backpropagation and training for neural networks [30]. These strategies, however, are time consuming and can scale poorly with circuit size. Moreover, it is often inefficient to retrain hardware settings for each individual chip. For many tasks, such as machine learning, model training is energy intensive; if the same model parameters are broadcast to thousands of chips within a data center, retraining the model for each chip with a unique set of component imprecisions will be very costly. One can instead employ progressive algorithms making use of local feedback [31,32]; however, these algorithms, which iteratively optimize the settings of one device at a time, require $O(N^2)$ tap photodiodes to monitor the optical power within each individual interferometer. This requirement greatly increases the number of electrical lines and overall power consumption of the system.

This focus on *in situ* approaches reveals a critical roadblock for programmable photonics compared to electronic FPGAs.

2334-2536/21/101247-09 Journal © 2021 Optica Publishing Group



Problem: calibration procedure (to measure errors) is really involved, requires very accurate measurements.

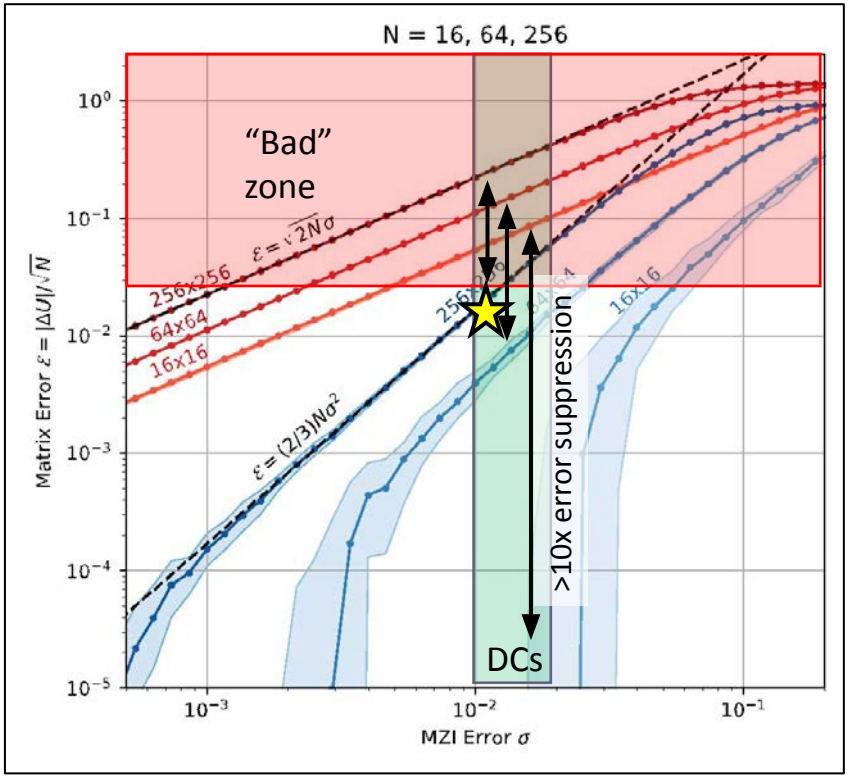
If errors are *known*, they can (mostly) be corrected by a low-cost deterministic procedure!

Benefits of Error Correction

Uncorrected error: $\epsilon_0 = \sqrt{2N}\sigma$
- Linear in σ (obviously)
- $\sqrt{N}$ dependence: *N* layers deep, errors add in quadrature

Corrected error: $\epsilon_c = (2/3)N\sigma^2$
- Quadratic in σ, better for small errors.
- Still increases with *N* (EC doesn't let us scale up MZI mesh arbitrarily)
- Quadratic error suppression: $\epsilon_c = (\epsilon_0)^2/3$

Cutoff threshold: $\sigma_{th} = \sqrt{3/N^3}$
- Perfect error correction possible for small circuits…
- …but gets very difficult as circuits scale up.



Slide credit: R Hamerly

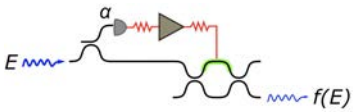
Example: Hardware EC for ONN Inference

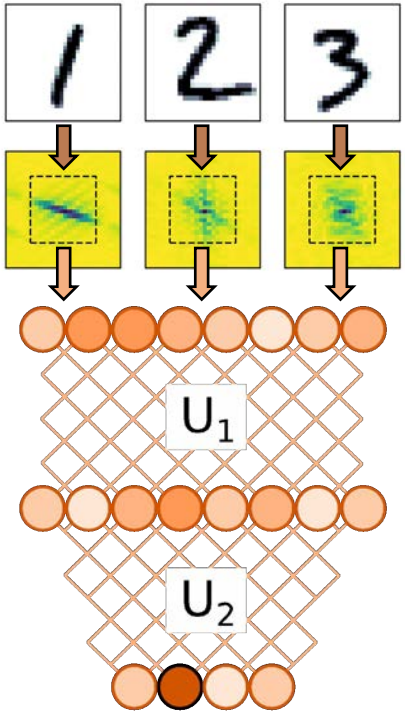
MNIST digit dataset

Preprocessing with FFT

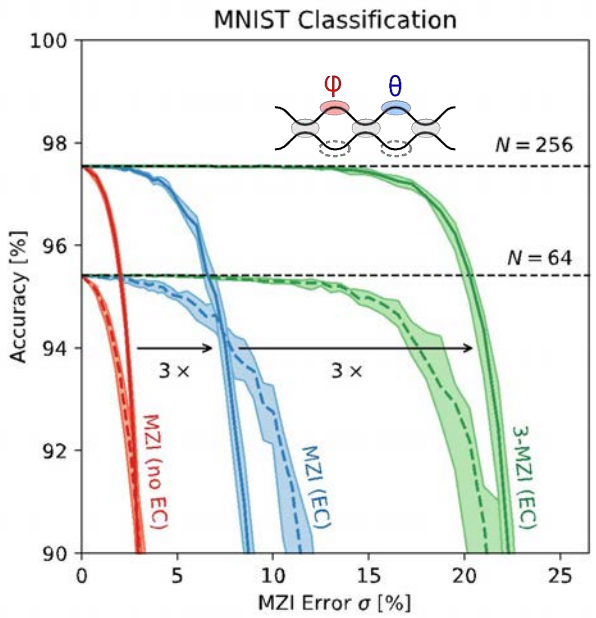
Filter to a $\sqrt{N}\times\sqrt{N}$ window.
N-neuron input to our ONN.

2-layer ONN.
$N\times N$ Clements meshes.
Electo-optic nonlinearity.





1, 2, 3, 4, ...

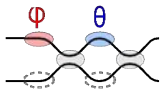
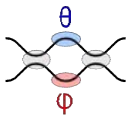
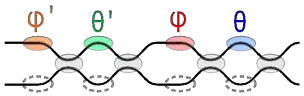
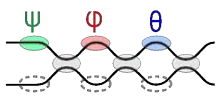
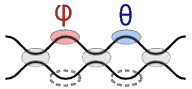
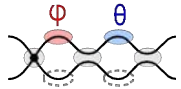


Based on: S. Pai et al., IEEE JSTQE 26, 6100813 (2020)
I. Williamson et al., IEEE JSTQE 26, 7700412 (2020)

Slide credit: R Hamerly

“Perfect” photonics from imperfect components?

yes with caveats*

		MZI	S-MZI [1]	Miller [2]	Suzuki [3]	**3-MZI**	**MZI+X**
							
Resources	Passives	2	2	4	3	3	3
	Actives	2	2	4	3	2	2
	Area	1.0	0.8	2.0	1.5	1.2	1.2
Features	Self-config	Y	-	Y	Y	Y	Y
	Broadband	-	-	-	-	-	Y
	Perfect	-	-	Y	Y	Y*	Y*†

S. Bandyopadhyay, *Optica* 8, 1247 (2021)

Ryan Hamerly, Saumil Bandyopadhyay, DE, “Asymptotically fault-tolerant programmable photonics” N COMM **13** (2022)

* In the limit $N \to \infty$ † Correlated errors only

[1] Bell & Walmsley, arXiv:2104.07561 [2] Miller, *Optica* **2**, 747 (2015) [3] Suzuki, *OpEx* **23**, 9086 (2015)

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing
- EC in hardware compute, training: → "error-free optics"
- EC in ML in-situ training?

Review: G Wetzstein et al, Nature 588 (2020)

Review: W Bogaerts et al, Nature 586, 207–216 (2020)

S. Bandyopadhyay, Optica 8, 1247 (2021)

R Hamerly et al, Nature Comm. 13 (2023)

Key point #3: Coherent error cancellation → **Fault tolerant matrix operations**

Applications

Article | Open Access | Published: 29 November 2022

Asymptotically fault-tolerant programmable photonics

Ryan Hamerly, Saumil Bandyopadhyay & Dirk Englund

Nature Communications **13**, Article number: 6831 (2022) | Cite this article

PHYSICAL REVIEW APPLIED **18,** 024019 (2022)

Accurate Self-Configuration of Rectangular Multiport Interferometers

Ryan Hamerly,[1,2,*] Saumil Bandyopadhyay,[1] and Dirk Englund[1]

[1]*Research Laboratory of Electronics, MIT, 50 Vassar Street, Cambridge, Massachusetts 02139, USA*

[2]*NTT Research Inc., PHI Laboratories, 940 Stewart Drive, Sunnyvale, California 94085, USA*

(Received 21 January 2022; revised 27 June 2022; accepted 28 June 2022; published 5 August 2022)

PHYSICAL REVIEW APPLIED **18,** 024018 (2022)

Stability of Self-Configuring Large Multiport Interferometers

Ryan Hamerly,[1,2,*] Saumil Bandyopadhyay,[1] and Dirk Englund[1]

[1]*Research Laboratory of Electronics, MIT, 50 Vassar Street, Cambridge, Massachusetts 02139, USA*

[2]*NTT Research Inc., PHI Laboratories, 940 Stewart Drive, Sunnyvale, California 94085, USA*

(Received 21 January 2022; revised 10 May 2022; accepted 28 June 2022; published 5 August 2022)

Research Article Vol. 8, No. 10 / October 2021 / Optica 1247

OPTICA

Hardware error correction for programmable photonics

SAUMIL BANDYOPADHYAY,[1,*] RYAN HAMERLY,[1,2] AND DIRK ENGLUND[1]

[1]Research Laboratory of Electronics, MIT, 50 Vassar Street, Cambridge, Massachusetts 02139, USA

[2]NTT Research Inc., PHI Laboratories, 940 Stewart Drive, Sunnyvale, California 94085, USA

*Corresponding author: saumilb@mit.edu

Received 2 March 2021; revised 1 July 2021; accepted 18 August 2021 (Doc. ID 424052); published 27 September 2021

883

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing
- EC in hardware compute, training: → "error-free optics"
- EC in ML in-situ training?
- Fully integrated on-chip inference

Review: G Wetzstein et al, Nature 588 (2020)

Review: W Bogaerts et al, Nature 586, 207–216 (2020)

S. Bandyopadhyay, Optica 8, 1247 (2021)

R Hamerly et al, Nature Comm. 13 (2023)

S Vadlamani, Science Advances (2023)

Applications

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] and Dirk Englund[1,†]

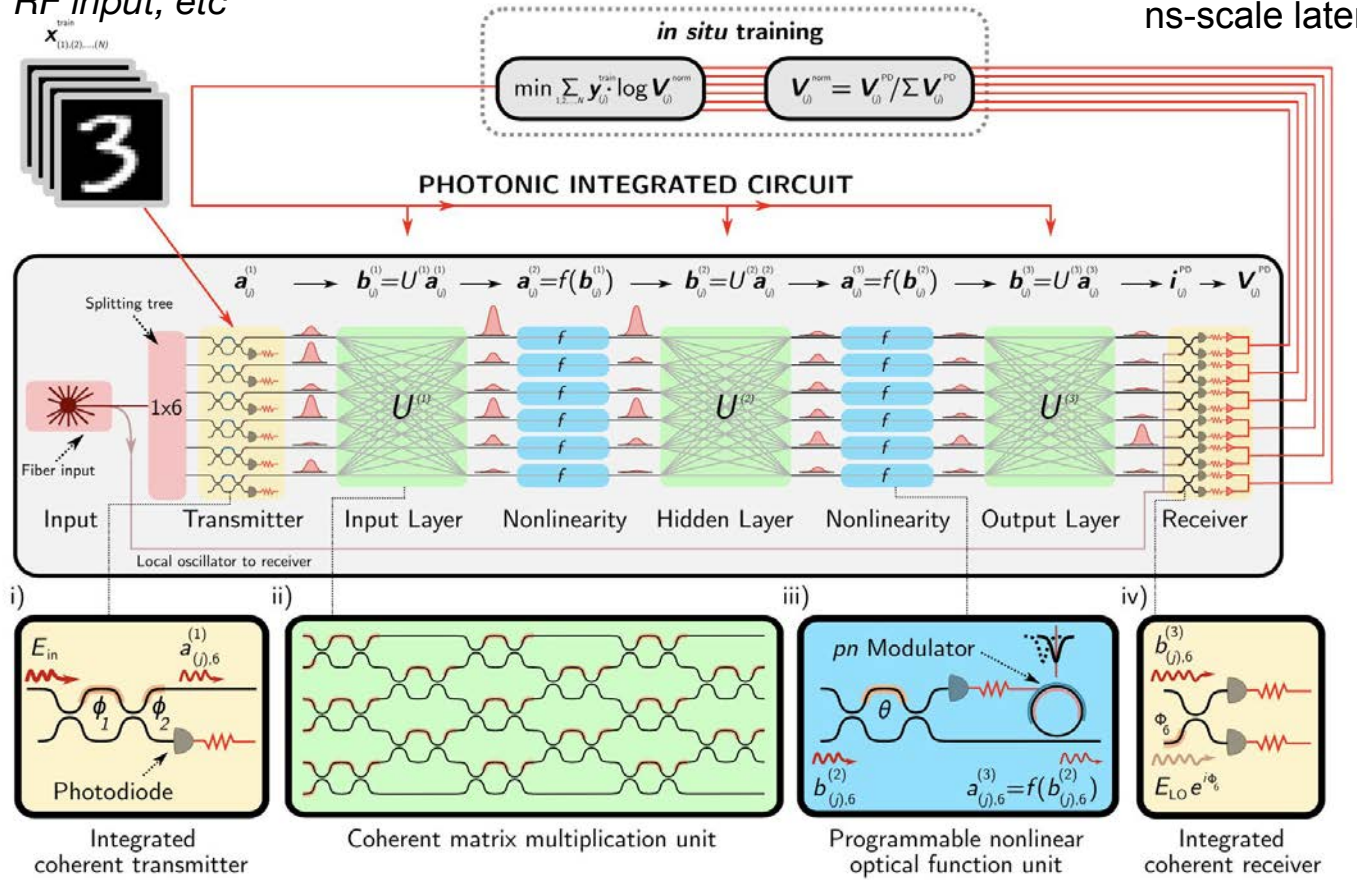
[1]*Research Laboratory of Electronics, MIT, Cambridge, MA 02139, USA*
[2]*NTT Research Inc., PHI Laboratories, 940 Stewart Drive, Sunnyvale, CA 94085, USA*
[3]*Nokia Corporation, New York, NY, 10016, USA*
[4]*Luminous Computing Inc., Mountain View, CA, 94041, USA*

Saumil Bandyopadhyay

RF-Photonic ML: O(M N^2) → O(M N) effort for $\mathbf{x_N} \rightarrow f_{DNN}(\mathbf{x_N})$ and M layers

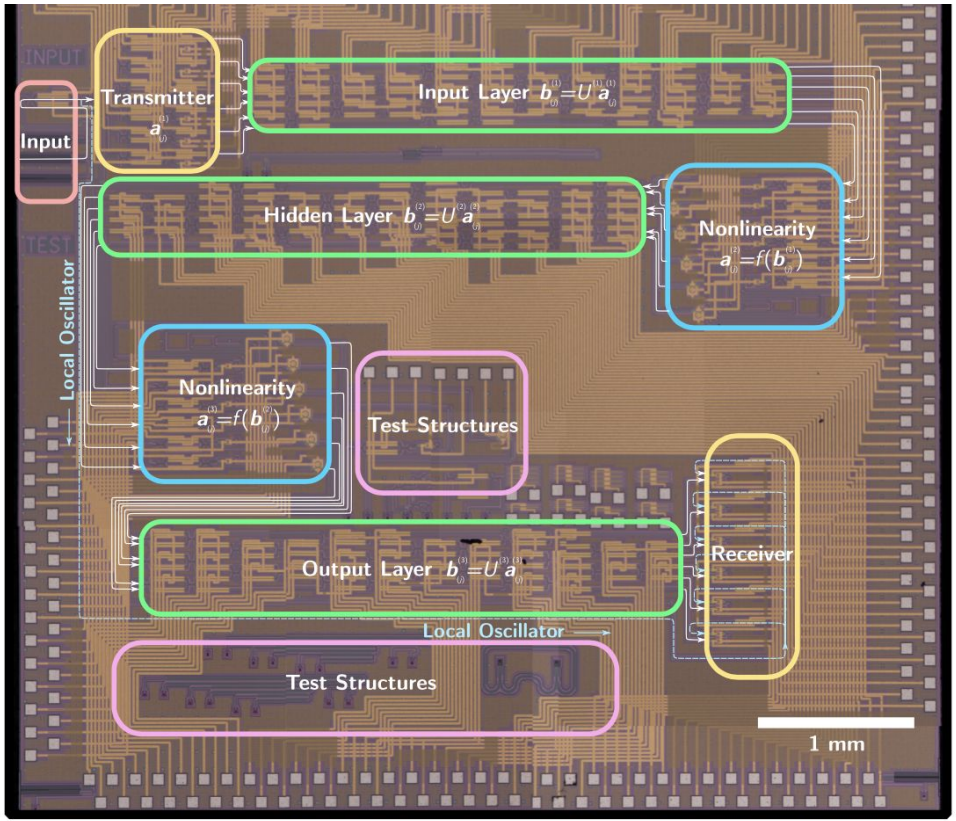
*input **x** from memory, RF input, etc*

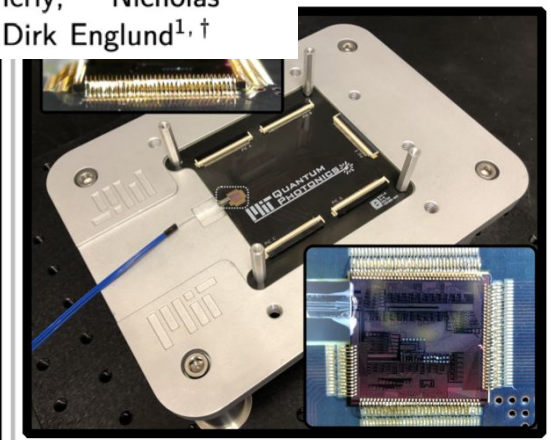
→ 10 fJ/OP for large systems and ns-scale latency

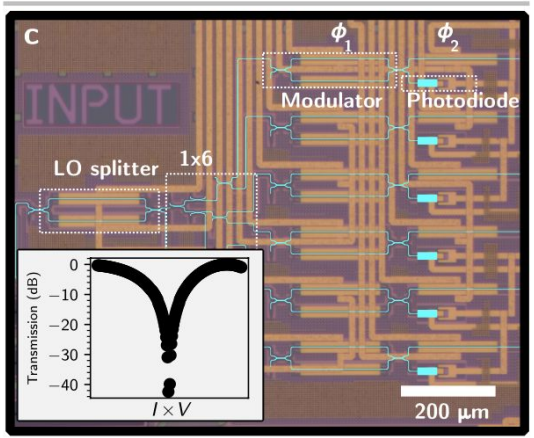


Single chip photonic deep neural network with accelerated training

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] and Dirk Englund[1,†]



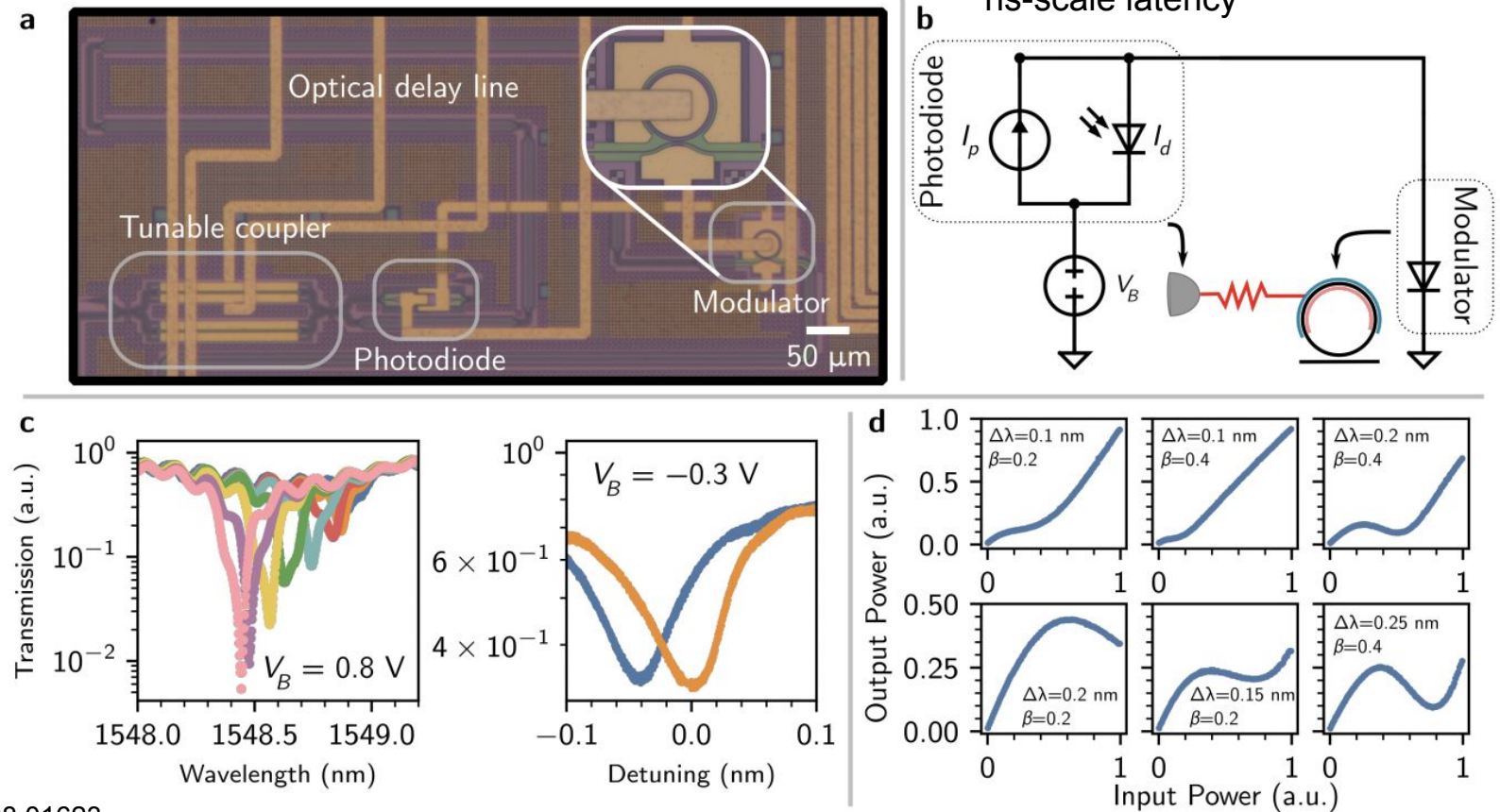




arXiv:2208.01623

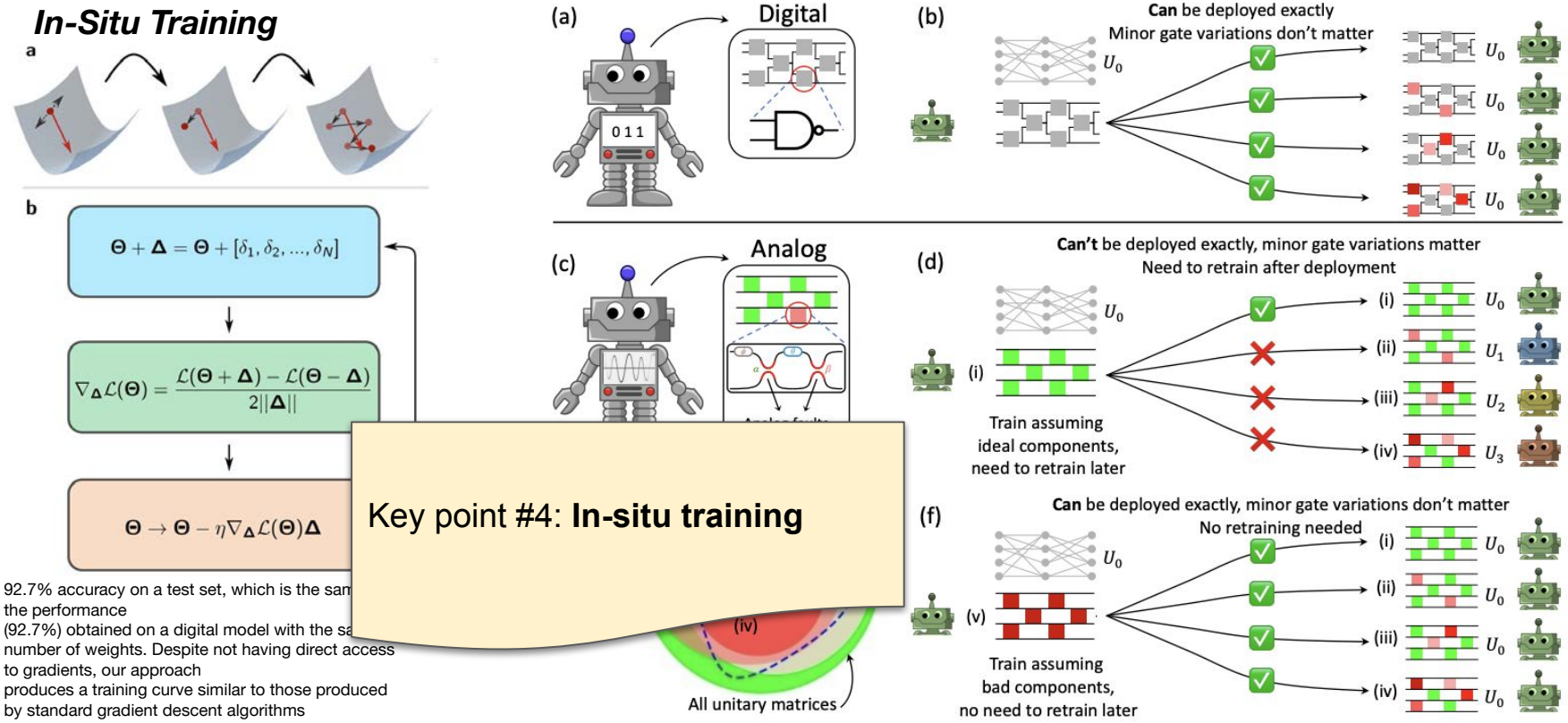
Nonlinear Optical Function Unit ("NOFU")

→ 10 fJ/OP for large systems and ns-scale latency



arXiv:2208.01623

In-situ training: better .. and transferable?



Key point #4: **In-situ training**

92.7% accuracy on a test set, which is the sam
the performance
(92.7%) obtained on a digital model with the sa
number of weights. Despite not having direct access
to gradients, our approach
produces a training curve similar to those produced
by standard gradient descent algorithms

S. Bandyopadhyay et al, arXiv:2208.01623 (2022)

Sri K Vadlamani,DE, R Hamerly, ``Transferable Learning on Analog Hardware,'' Science Advances (2022)

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing

Review: G Wetzstein et al, Nature 588 (2020)

Review: W Bogaerts et al, Nature 586, 207–216 (2020)

- EC in hardware compute, training: → "error-free optics"

S. Bandyopadhyay, Optica 8, 1247 (2021)

R Hamerly et al, Nature Comm. 13 (2023)

- EC in ML in-situ training?

S Vadlamani, Science Advances (2023)

- Fully integrated on-chip inference

Applications

- Computing Across the Internet's Edge: *Tops/sec @mW on edge devices*

S. Bandyopadhyay, *arXiv:2203.05466* (2022)

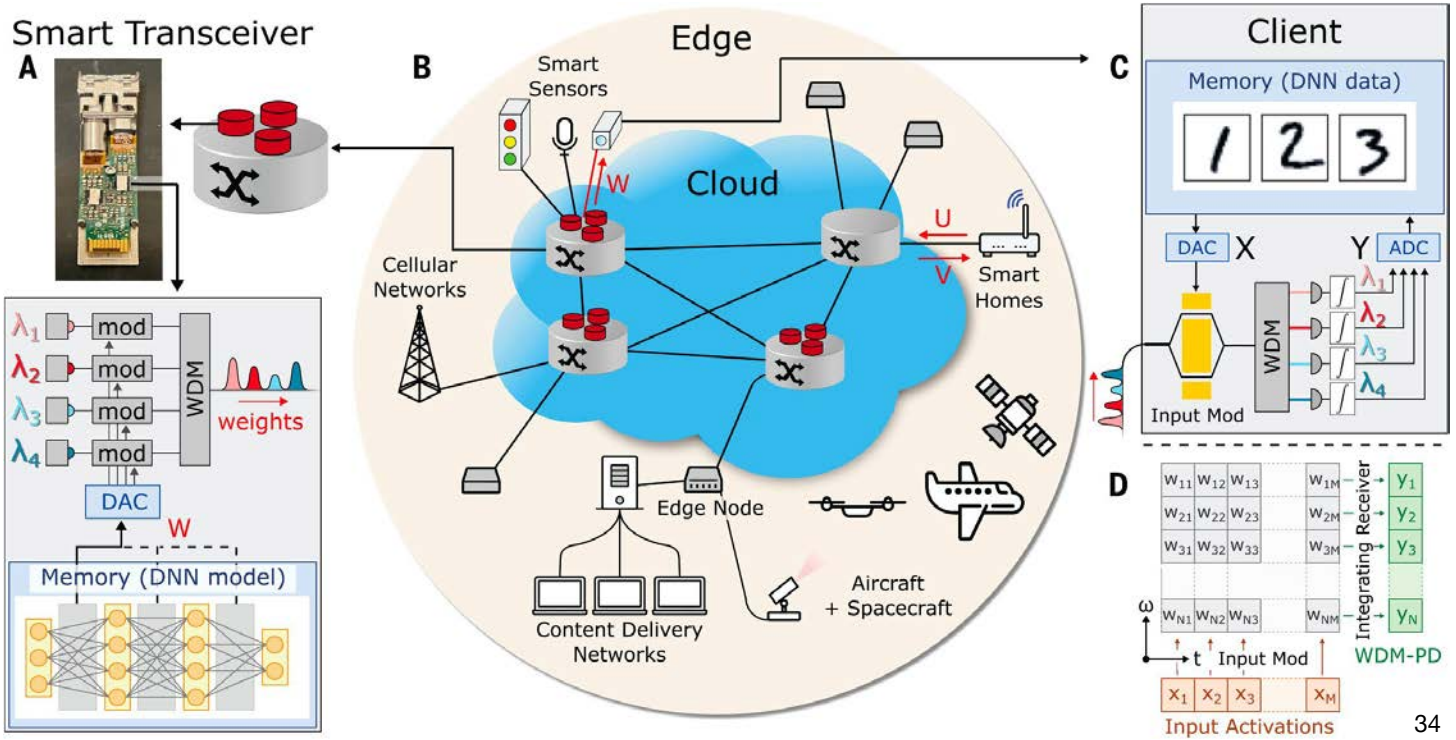
6883

Hey Siri

COMPUTER NETWORKS

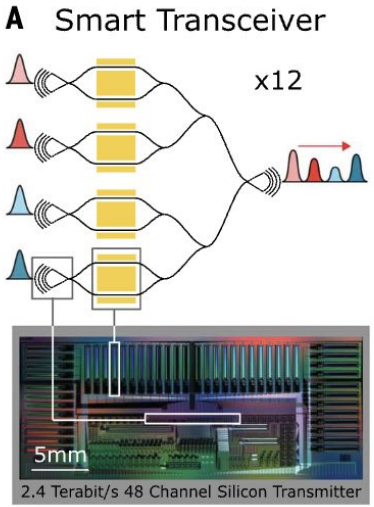
Delocalized photonic deep learning on the internet's edge

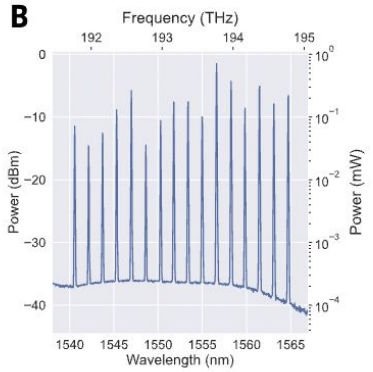
Alexander Sludds[1]*, Saumil Bandyopadhyay[1], Zaijun Chen[1]†, Zhizhen Zhong[2], Jared Cochrane[1,3], Liane Bernstein[1], Darius Bunandar[1]‡, P. Ben Dixon[3], Scott A. Hamilton[3], Matthew Streshinsky[4]§, Ari Novack[4]§, Tom Baehr-Jones[4]§, Michael Hochberg[4]§, Manya Ghobadi[2], Ryan Hamerly[1,5]*,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

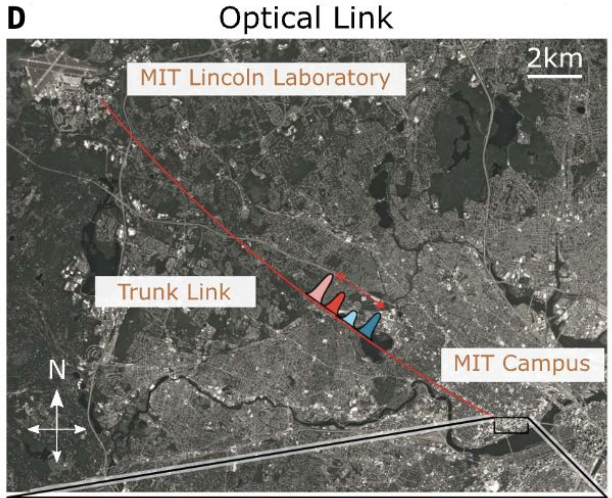


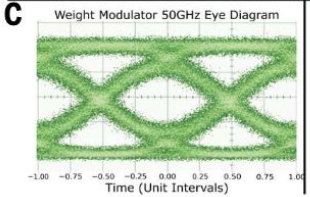
Sludds et al.,Science 378 (2022)

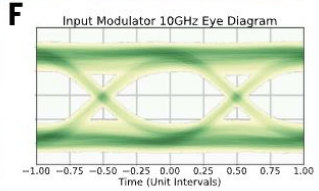
Experiment

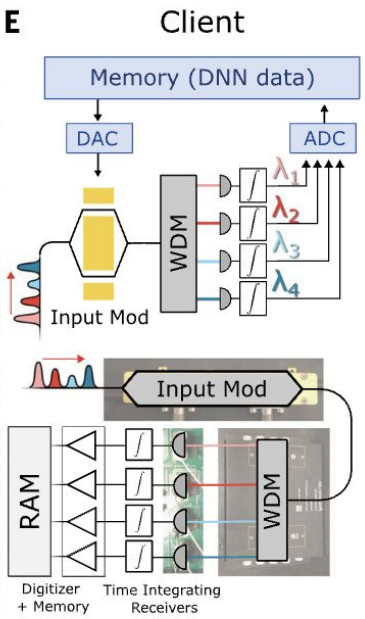


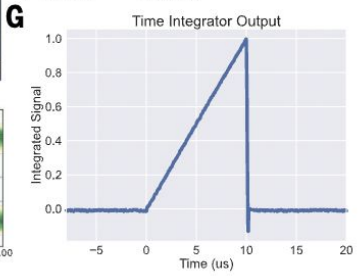












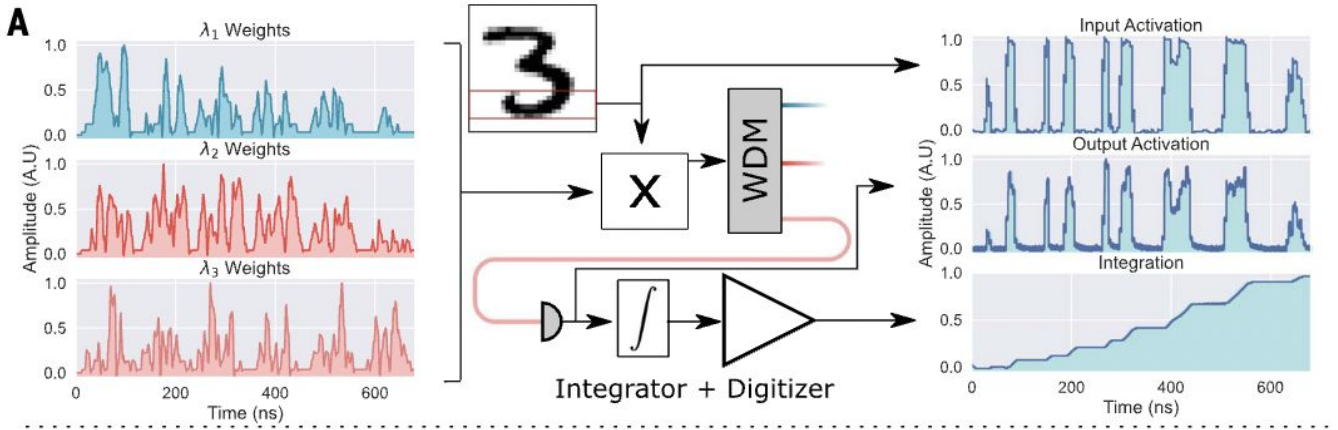
We use a 2.4 Tbps silicon smart transceiver. Made in a commercial CMOS foundry

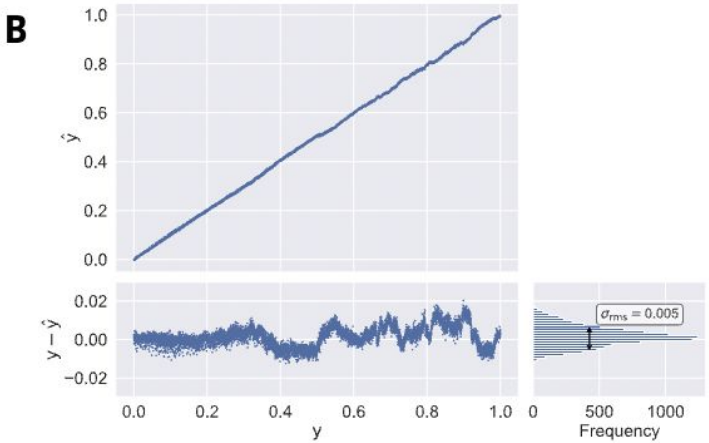
Deploy weights over 86 km deployed fiber in a Boston area fiber network.

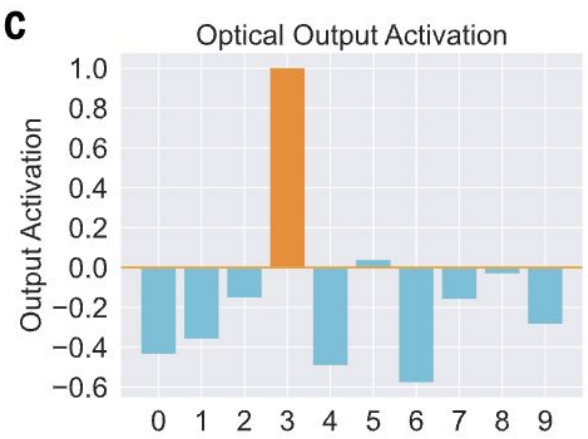
Lightweight edge clients are able to compute with ~mW of power at THz rates with one GHz modulator and detecting at MHz rates.

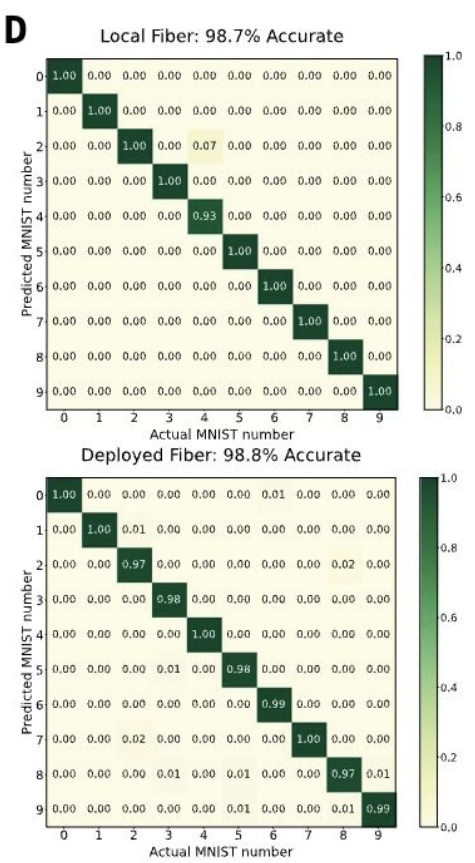
Sludds et al.,Science 378 (2022)

Accuracy on MNIST (28x28)









Sludds et al.,Science 378 (2022)

SNR essentials

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

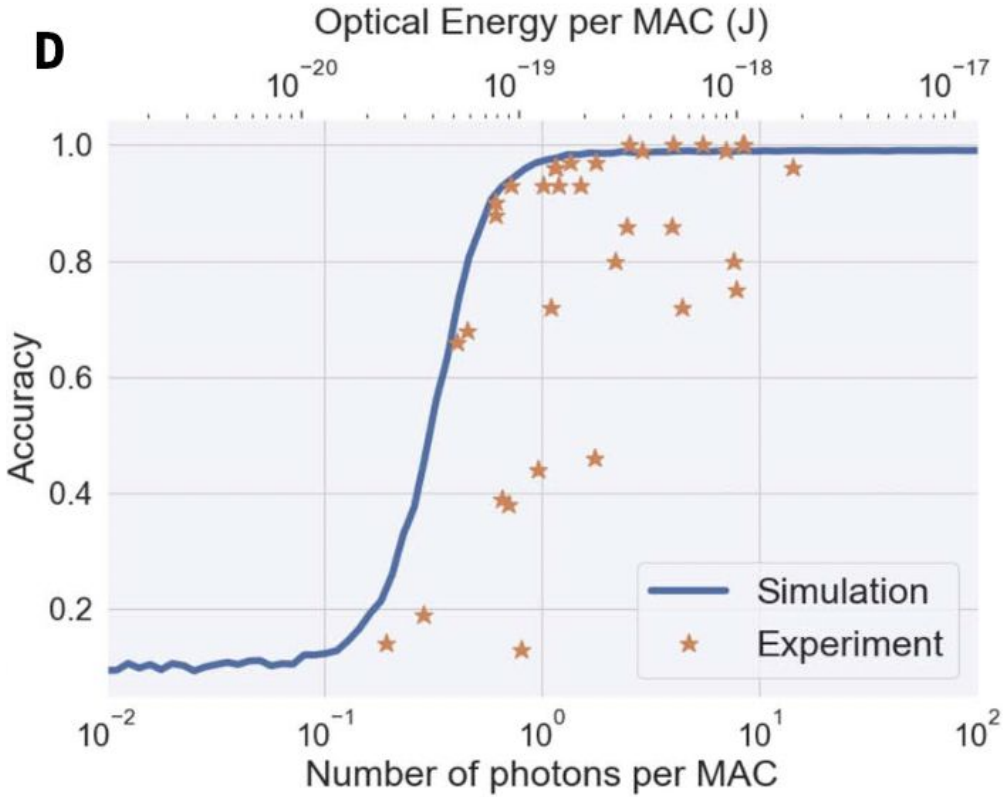
Table 1. Device contributions to receiver performance assuming conventional technology. Device energy consumption is amortized by either a spatial fan-out factor (*N*) or time-domain fan-out factor (*M*). We assume a carrier depletion modulator in silicon is used and that a single high-speed (gigahertz) 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

Sludds et al.,Science 378 (2022)

Near-perfect inference with ~1 photon/symbol

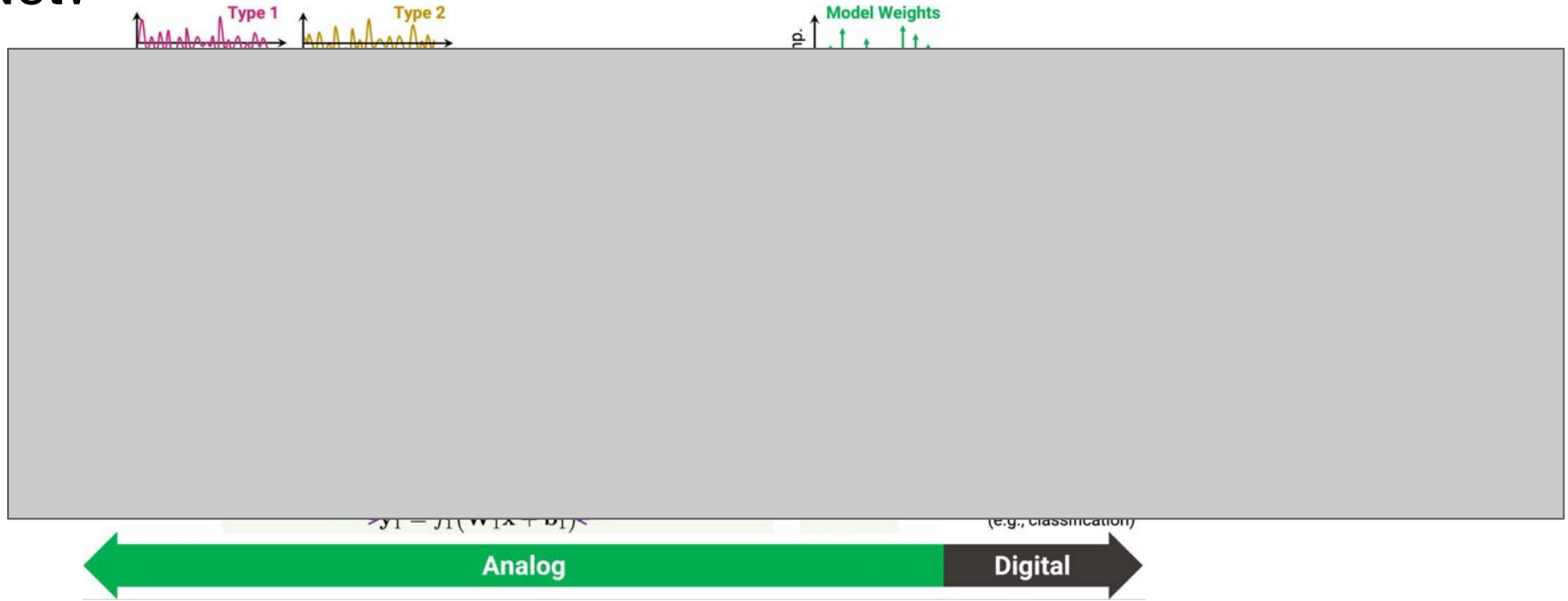


Sludds et al.,Science 378 (2022)

Recent extension to ultra-low-power radios

Machine Intelligence on Wireless Edge Networks (MIWEN)

Collaboration with Prof Tingjun Chen (Duke)



Contact: englund@mit.edu, tingjun.chen@duke.edu (cc erikyost@mit.edu)

Recent extension to ultra-low-power radios

Machine Intelligence on Wireless Edge Networks (MIWEN)

Collaboration with Prof Tingjun Chen (Duke)

- Disaggregated memory access and ML inference in wireless networks
- Compute moved into existing RF/analog chain of wireless signal processing
- “Smart-TRXs” eliminate or sharply reduce A/D and E/O conversions
- Sharp cut in power consumption and latency (S. Vadlemani et al; ..)
- New physical-layer security and privacy (Kfir Salon et al, in preparation)
-
- Energy consumption for ML inference will be below the thermodynamic bound of digital electronics

Contact: englund@mit.edu, tingjun.chen@duke.edu (cc erikyost@mit.edu)

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing

Review: G Wetzstein et al, Nature 588 (2020)

Review: W Bogaerts et al, Nature 586, 207–216 (2020)

- EC in hardware compute, training: → "error-free optics"

S. Bandyopadhyay, Optica 8, 1247 (2021)

R Hamerly et al, Nature Comm. 13 (2023)

- EC in ML in-situ training?

S Vadlamani, Science Advances (2023)

- Fully integrated on-chip inference

Applications

- Computing Across the Internet's Edge: *Tops/sec @mW on edge devices*

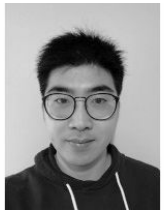
S. Bandyopadhyay, *arXiv:2203.05466* (2022)

- Microlaser-based Coherent Scalable Efficient deep Learning *~ fJ/OP and 25 TeraOP/(mm^2· s)*

A. Sludds et al, Science (2022)

Z. Chen et al. Nature Photonics (2023)

883



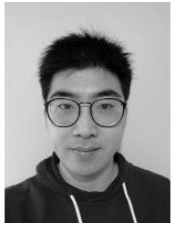
Dr. Zaijun Chen (→ USC)

SPIE Photonics west best paper award
Z Chen et al, Nature Photonics (2023)
2024 Optica Challenge Award winner

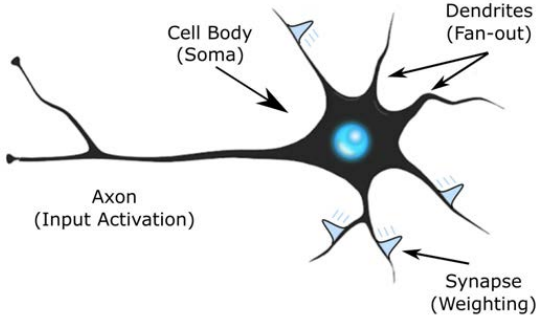
ONNs in networks of VCSELs

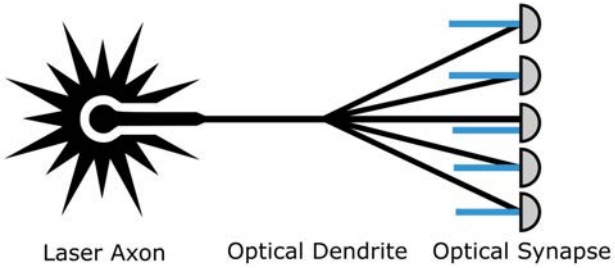
Output stationary ONN with photoelectric multiplication

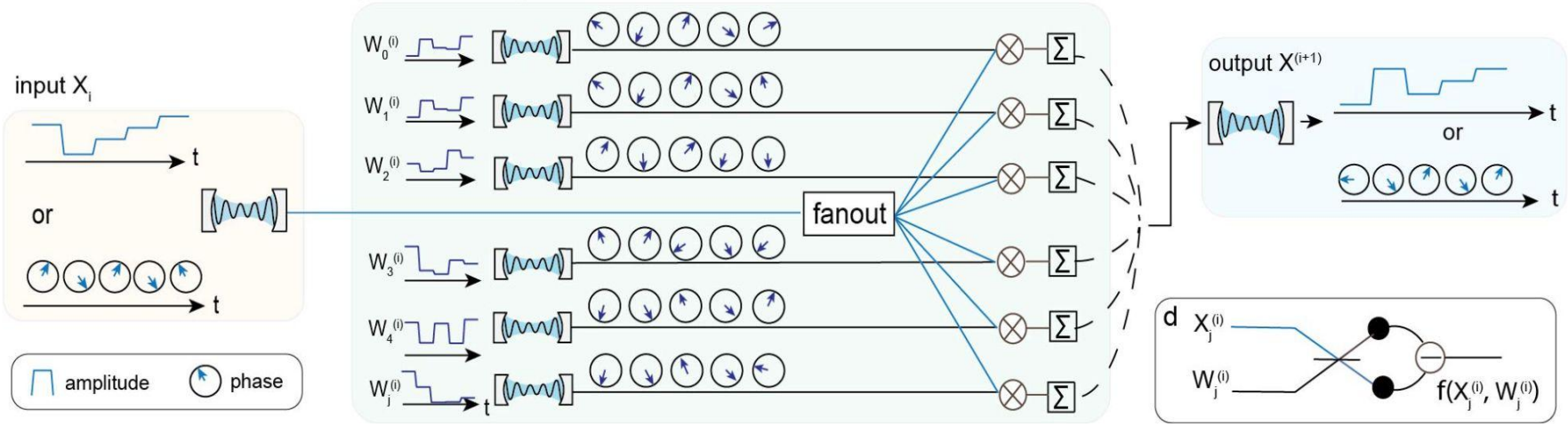
With Prof Stephan Reitzenstein
@TU-Berlin



Zaijun Chen
(--> USC)



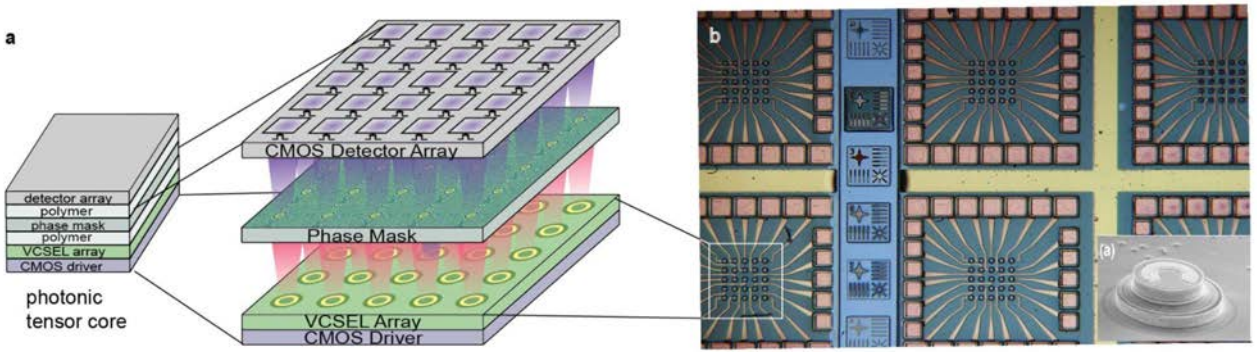




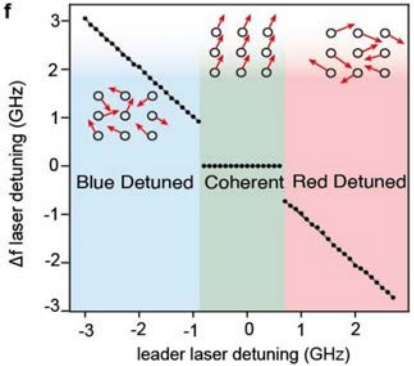
Z. Chen et al, Photonics (2023) - Theory based on photoelectric multiplication (PRX 9, 021032 (2019))

Compute density, energy efficiency, nonlinearity

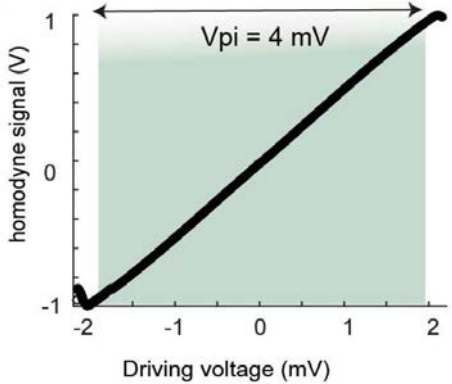
3D architecture



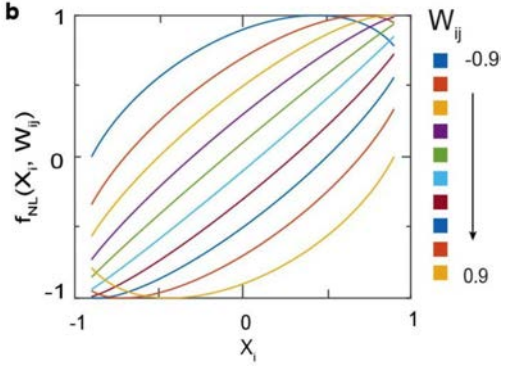
Coherent injection lock



Ultralow Vπ = 4mV

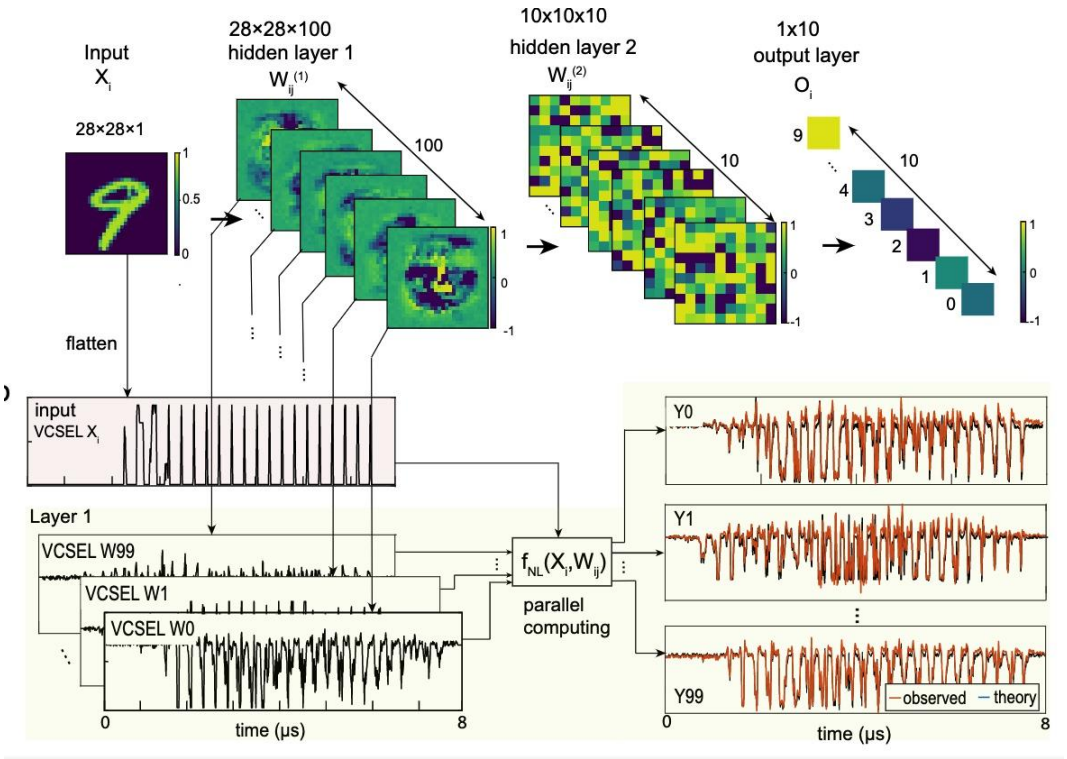


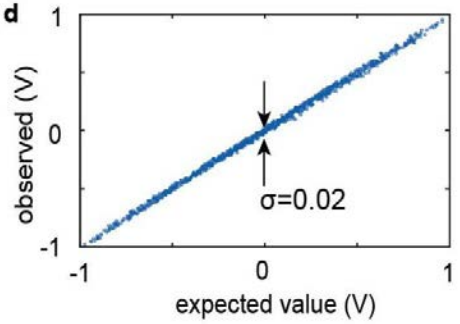
Detection-based nonlinearity

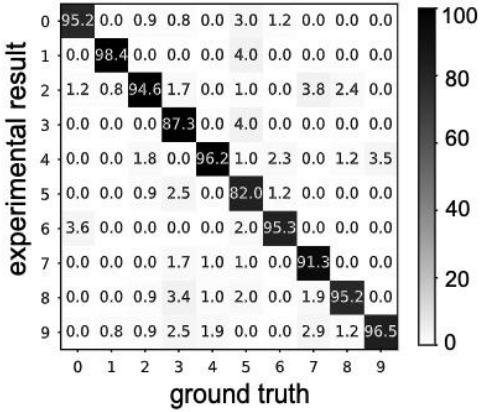


(1) coherent on-chip microscale VCSEL transmitters as high-speed (GHz) ‘laser axons’,
(2) coherent detection for weighted accumulation as low-energy ‘laser synapses’,
(3) holographic data movement as optical dendritic fanout
(4) detection-based optical nonlinearity

Large-scale parallel processing







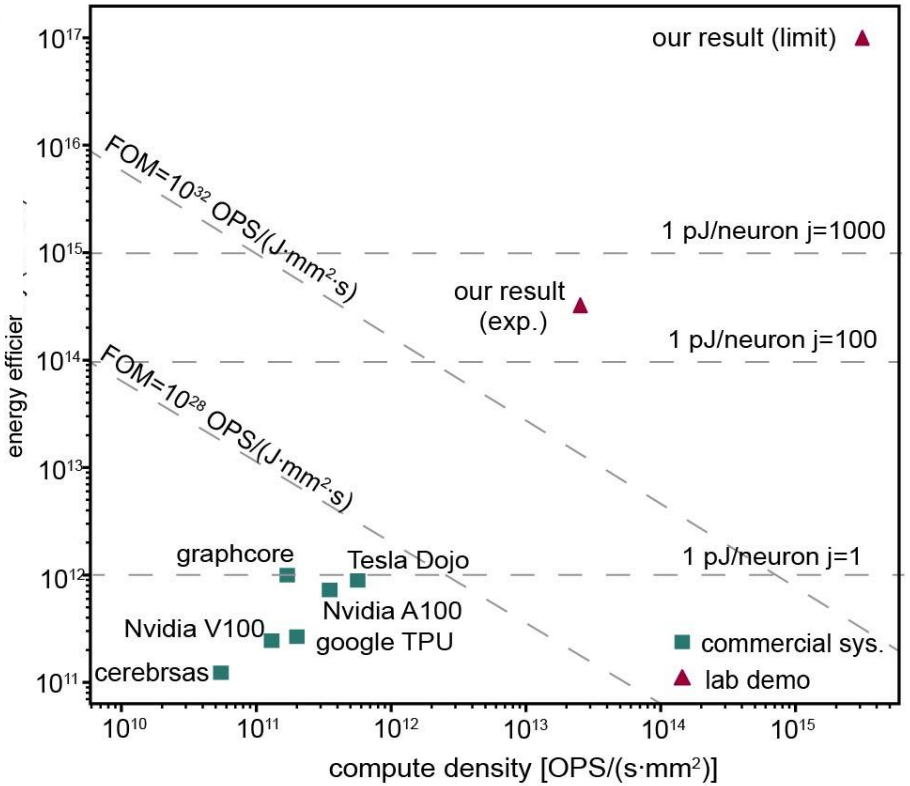
(1) 6 bits precision
(2) 93% classification accuracy
(3) Parallel computing with multi-channels

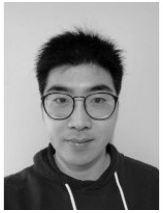
Obtained from inference over a random dataset of 1000 images, with a total of 158.8-million operations → accuracy of (93.1 ± 2)% vs model accuracy in simulation (95.1%).

Summary

Table 1 figures of merit of machine learning hardware

criteria	description	VCSEL ONN
C1	compute density ρ	25 TOP/(mm$^2 \cdot s$)
C2	energy consumption ϵ	2.5 fJ/OP
C3	inline nonlinearity	instantaneous response
C4	hardware scalability	wafer-scale volume production
C5	number of neurons	25 billion neurons per second





Dr. Zaijun Chen (→ USC)

SPIE Photonics west best paper award
Z Chen et al, Nature Photonics (2023)
2023 Optica Challenge Award winner

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing

Review: G Wetzstein et al, Nature 588 (2020)

Review: W Bogaerts et al, Nature 586, 207–216 (2020)

- EC in hardware compute, training: → "error-free optics"

S. Bandyopadhyay, Optica 8, 1247 (2021)

R Hamerly et al, Nature Comm. 13 (2023)

- EC in ML in-situ training?

S Vadlamani, Science Advances (2023)

- Fully integrated on-chip inference

Applications

- Computing Across the Internet's Edge: *Tops/sec @mW on edge devices*

S. Bandyopadhyay, *arXiv:2203.05466* (2022)

- Microlaser-based Coherent Scalable Efficient deep Learning *~ fJ/OP and 25 TeraOP/(mm^2· s)*

A. Sludds et al, Science (2022)

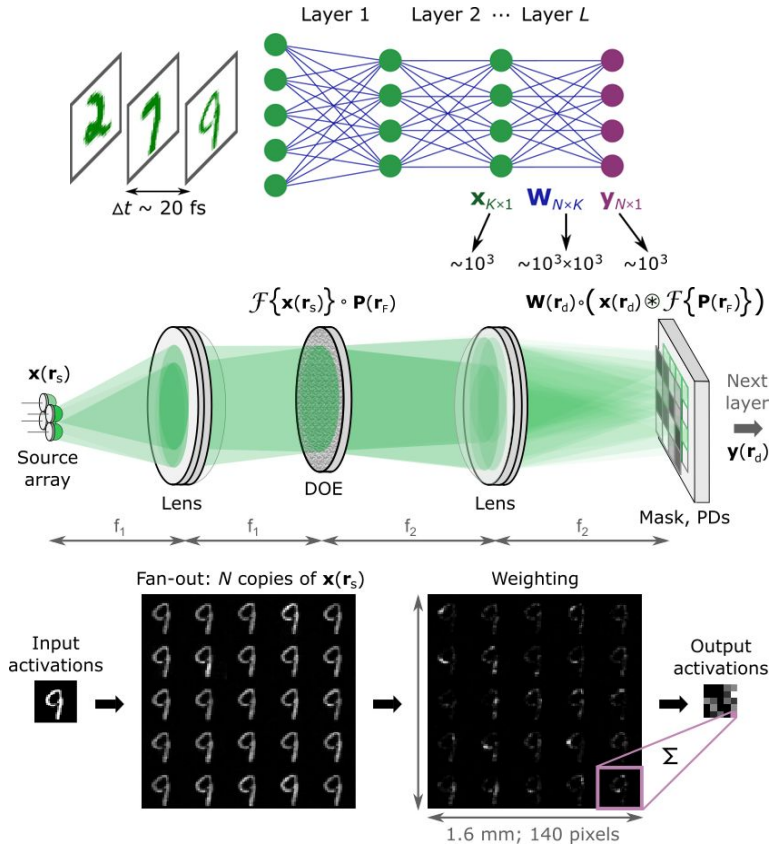
- Single-shot NNs: *20,000 MACs / 20 fs: Exascale computing*

Z. Chen et al, Nature Photonics (2023)

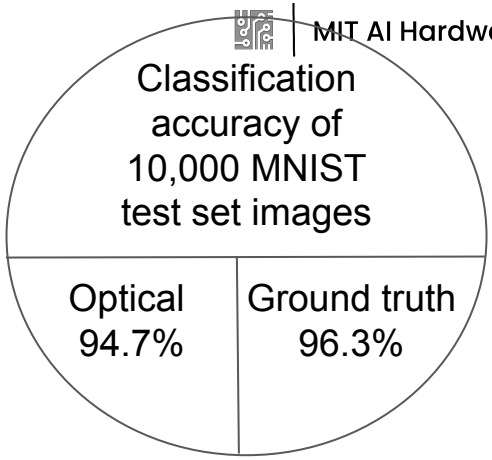
see also Ronald Davis III et al, ArXiv:2207.06883

L Bernstein et al, Science Advances

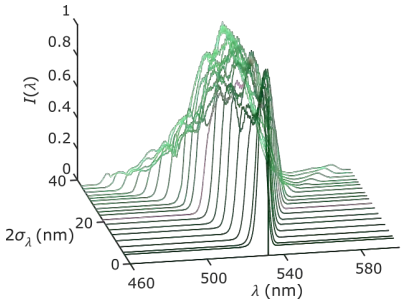
Single-shot ONN

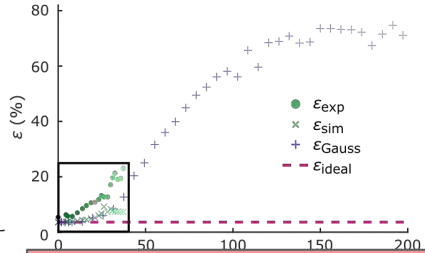


Liane Bernstein et al, Science Advances (2023)



Liane Bernstein

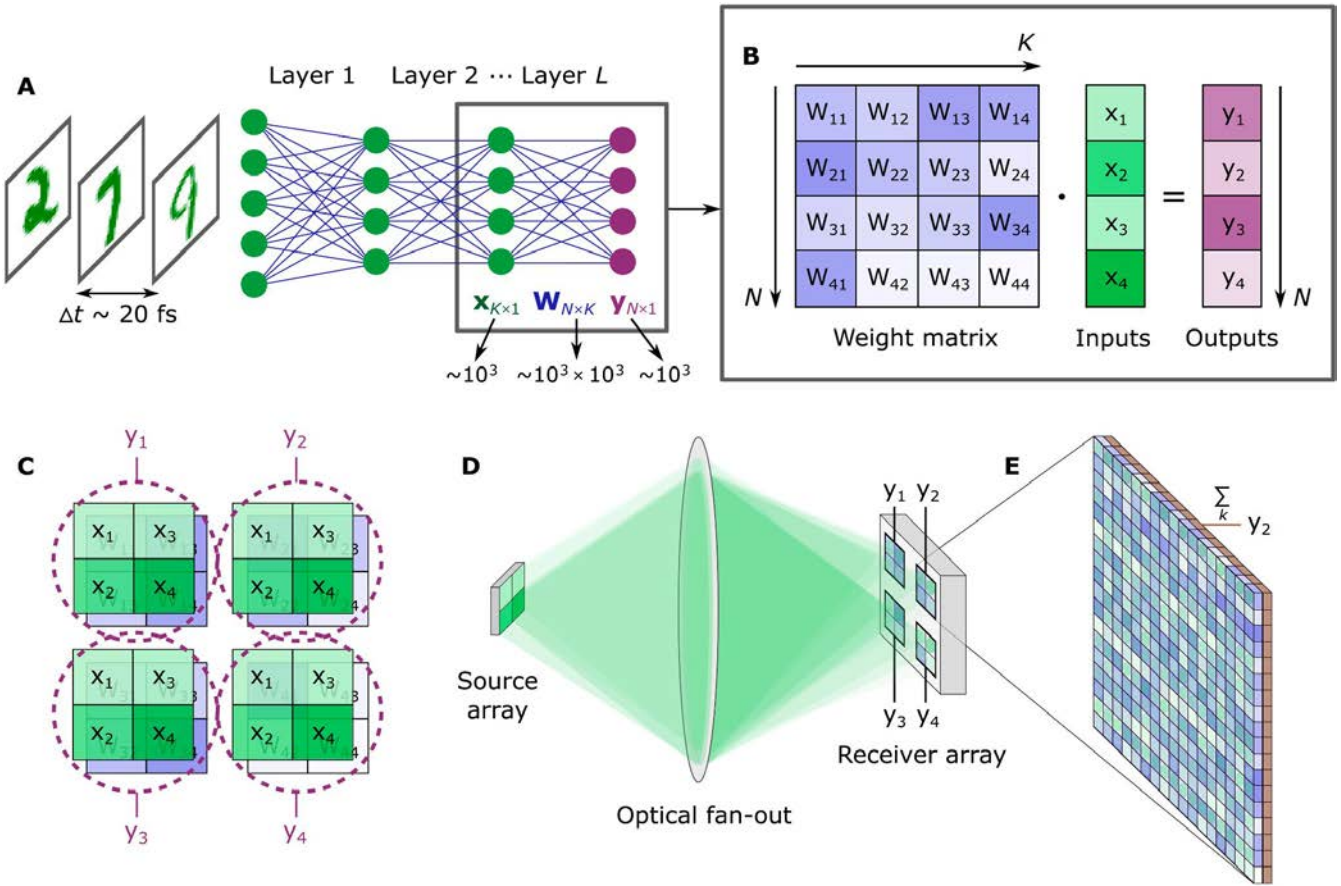




Theoretically the optics
w/ ~20fs pulses
20,000 MACs / 20 f

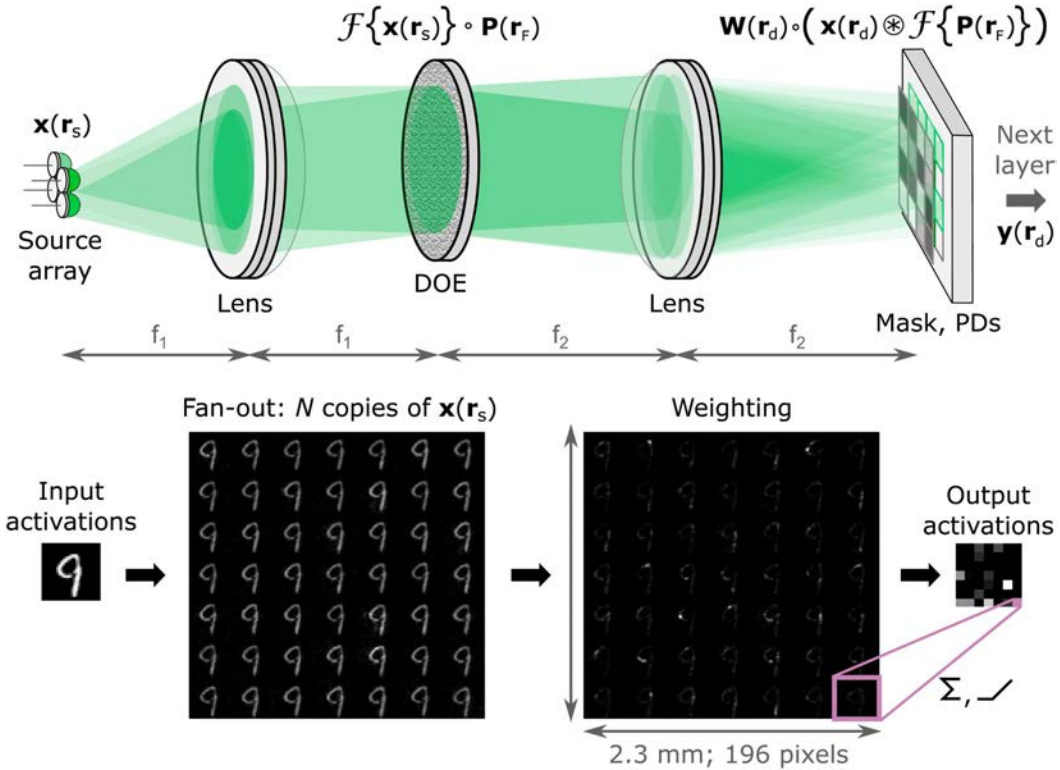
.. but hold on! that's a single query, but how many queries per second?! For this we need fast 2D transmitters..

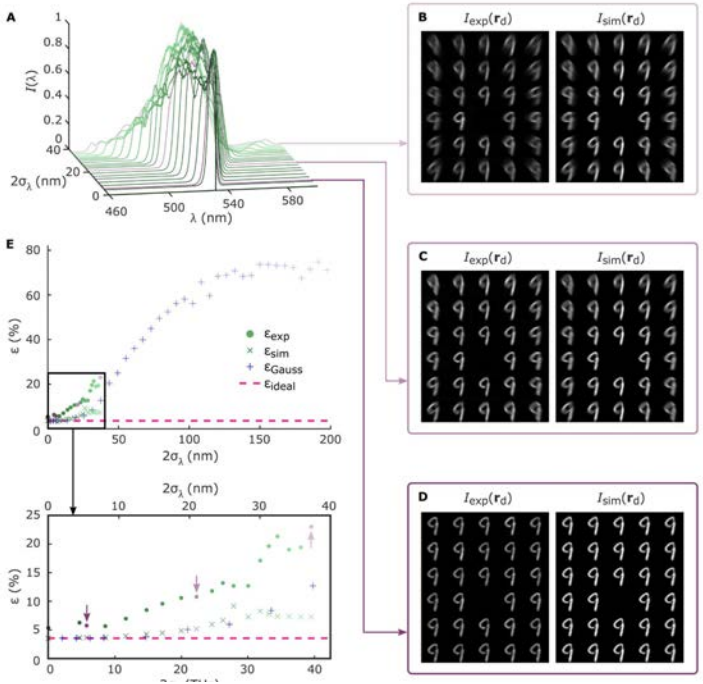
Single-shot deep neural network



Liane Bernstein

Experiment


$\mathcal{F}\{\mathbf{x}(\mathbf{r}_s)\} \circ \mathbf{P}(\mathbf{r}_F)$
$\mathbf{W}(\mathbf{r}_d) \circ (\mathbf{x}(\mathbf{r}_d) \circledast \mathcal{F}\{\mathbf{P}(\mathbf{r}_F)\})$
$\mathbf{x}(\mathbf{r}_s)$
Source array
Lens
DOE
Lens
Mask, PDs
Next layer
$\mathbf{y}(\mathbf{r}_d)$
f_1
f_1
f_2
f_2
Fan-out: N copies of $\mathbf{x}(\mathbf{r}_s)$
Weighting
Input activations
Output activations
2.3 mm; 196 pixels

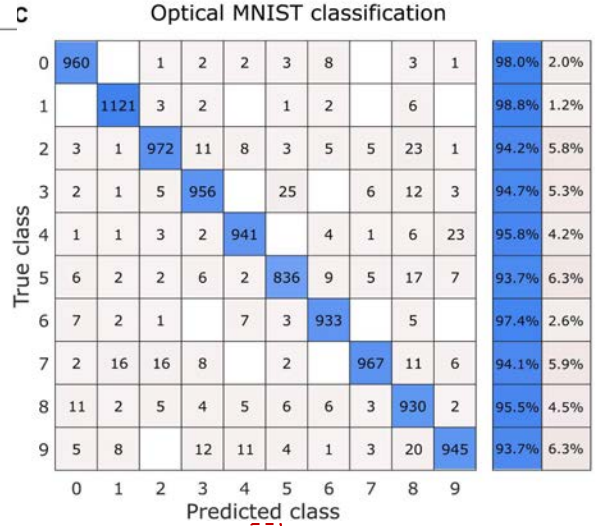

A
$I(\lambda)$
$2\sigma_\lambda$ (nm)
λ (nm)
B
$I_{exp}(\mathbf{r}_d)$
$I_{sim}(\mathbf{r}_d)$
C
$I_{exp}(\mathbf{r}_d)$
$I_{sim}(\mathbf{r}_d)$
D
$I_{exp}(\mathbf{r}_d)$
$I_{sim}(\mathbf{r}_d)$
E
ϵ (%)
ϵ_{exp}
ϵ_{sim}
ϵ_{Gauss}
ϵ_{ideal}
$2\sigma_\lambda$ (nm)

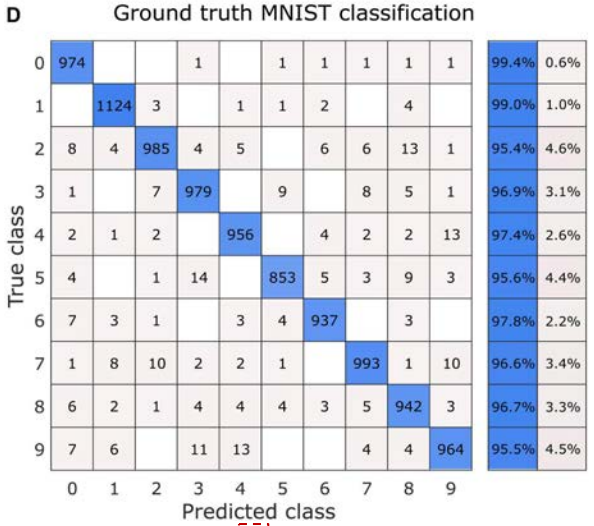
SCIENCE ADVANCES | RESEARCH ARTICLE

OPTICS

Single-shot optical neural network

Liane Bernstein[1]*, Alexander Sludds[1,2], Christopher Panuski[1], Sivan Trajtenberg-Mills[1], Ryan Hamerly[1,3], Dirk Englund[1]





Dataset	# Hidden layers	# Activations in hidden layer(s)	ONN accuracy	Ground-truth accuracy
MNIST	1	49	95.6	97.1
MNIST	2	25	95.1	96.4
Fashion-MNIST	2	36	83.3	87.1
QuickDraw*	2	36	79.0	82.6

Dataflow	Latency* (ns)	Throughput*,†
Weight stationary (near term)‡	~10	$N \times K$/1 ns→ ~petaMAC/s
Weight stationary (physical limit)§	~10	19,600/0.02 ps → ~exaMAC/s
Systolic array [e.g., TPU (7)]¶	$N + K \rightarrow 2000$	$N \times K$/1 ns→ ~petaMAC/s
Output stationary [e.g., (25)]	$K \rightarrow 1000$	$N \times K$/1 ns→ ~petaMAC/s

Table 2. Latency and throughput scaling of different architectures for computation of one DNN layer.
*Assuming $N = K = 1000$.†Assuming 100% utilization.‡Our proposed near-term optimized single-shot ONN (CMOS-compatible).§From maximum throughput measurements in Results.¶In practice, $N = K = 256$ (throughput ~100 teraMAC/s) in the TPU due to wiring and utilization constraints (7).

Estimates for arrays of high-speed transmitters

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing

Review: G Wetzstein et al, Nature 588 (2020)

Review: W Bogaerts et al, Nature 586, 207–216 (2020)

- EC in hardware compute, training: → "error-free optics"

S. Bandyopadhyay, Optica 8, 1247 (2021)

R Hamerly et al, Nature Comm. 13 (2023)

- EC in ML in-situ training?

S Vadlamani, Science Advances (2023)

- Fully integrated on-chip inference

Applications

- Computing Across the Internet's Edge: *Tops/sec @mW on edge devices*

S. Bandyopadhyay, *arXiv:2203.05466* (2022)

- Microlaser-based Coherent Scalable Efficient deep Learning *~ fJ/OP and 25 TeraOP/(mm^2· s)*

A. Sludds et al, Science (2022)

- Single-shot NNs: *20,000 MACs / 20 fs: Exascale computing*

Z. Chen et al, Nature Photonics (2023)

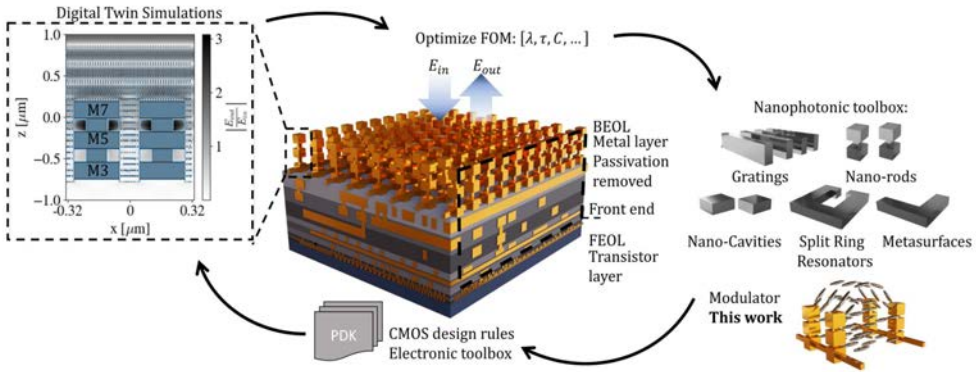
see also Ronald Davis III et al, ArXiv:2207.06883

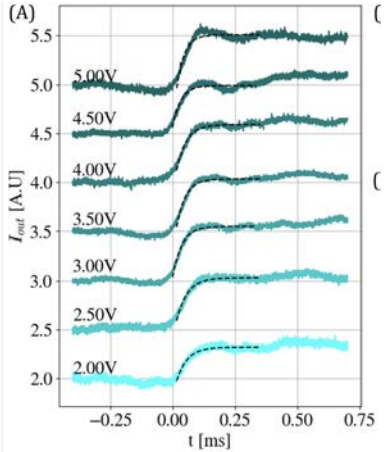
- New Tech for High-Speed Spatial Light Modulation

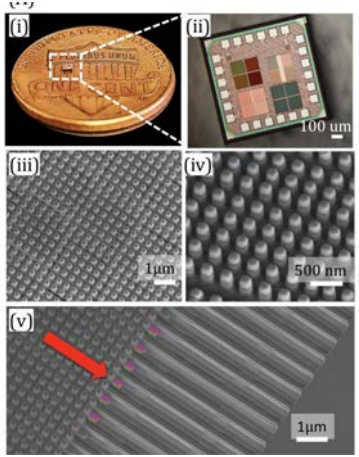
L Bernstein et al, Science Advances (2023)

M Dong et al, Nature Photonics (2021); C Panuski et al, Nature Photonics (2023); others in-review (see ArXiv.org)

High-speed SLMs + CMOS

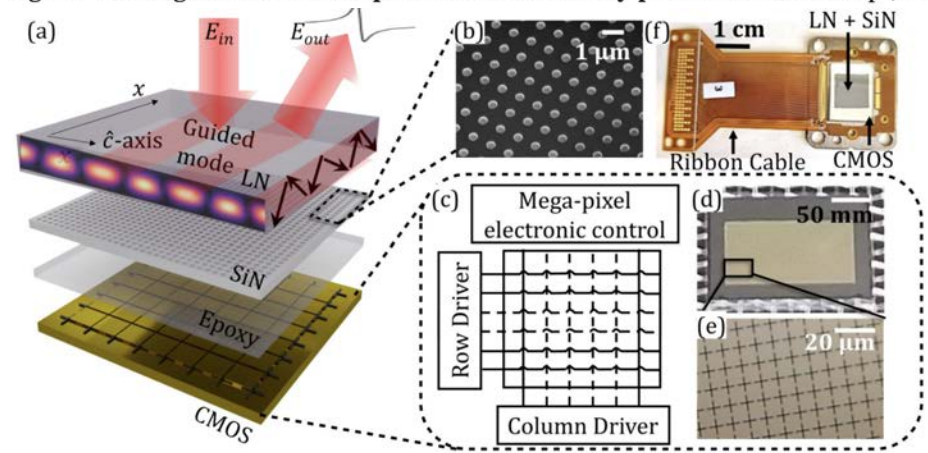


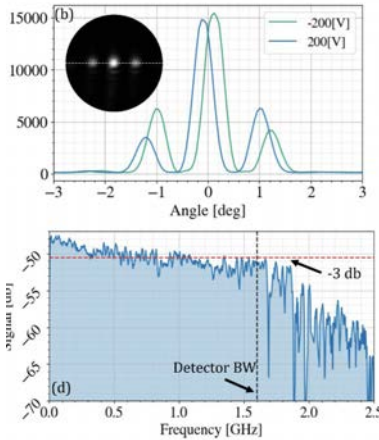




Mohamed ElKabbash, Sivan Trajtenberg-Mills, et al, physics>arXiv:2310.04409 (2023)

Fig. 1. Co-integration of Metal-optics in CMOS foundry processes. CMOS chips, com-



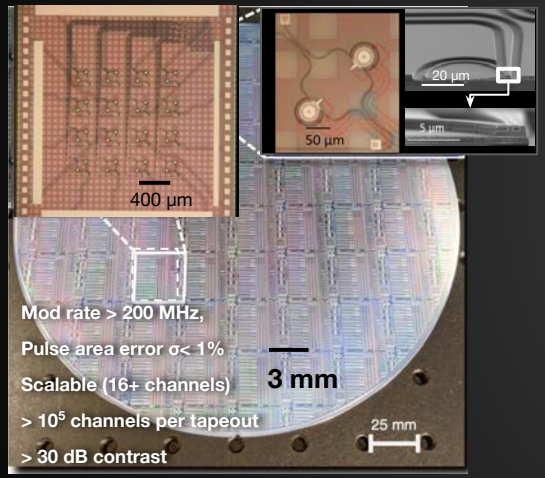


Sivan Trajtenberg-Mills,Mohamed ElKabbash, et al, "LNoS: Lithium Niobate on Silicon Spatial Light Modulator" ArXiV: 2402.1460 (2024)

2D Arrays of Modulators → High-Speed SLM (HS-SLM)

→ full-system demo in development w TSMC, DARPA

16-CH APIC[2] in 200 mm nanomechanical PIC (NEM-PIC) process [1]



[1] Dong et. al, Nature Photonics (2021)
[2]Adrian Menssen, Artur Hermans, et. al. Optica (2023)

MIT-Sandia National Laboratory - MITRE - Harvard
Separately: QuEra - SNL

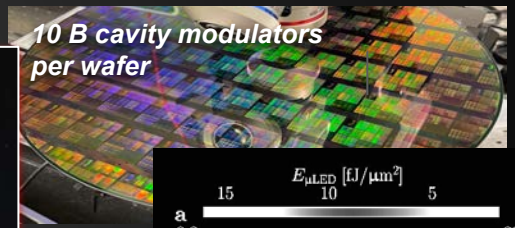
Process pioneered by Matt Eichenfield et al at Sandia NL

16-CH APIC lithium niobate on insulator[3]

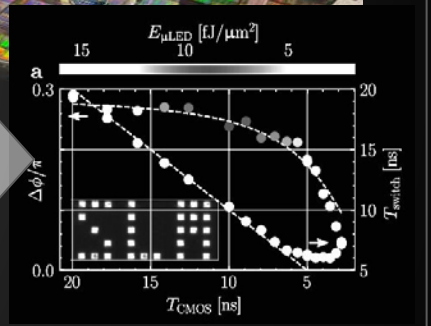
MIT QP csem

2D Spatial Light Modulator in Si demonstrator (v_{mod}~1GHz) [4]

AIM PHOTONICS



ns frames



C. Panuski et al, Nature Photonics (2022)
CP awarded the 2022 Carl E. Anderson Division of Laser cience Dissertation Award]

48-ch EOMs (v_{mod}>30 GHz):
A. Sludds et al, *Science* ***378*** (2022).

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing
- EC in hardware compute, training: → "error-free optics"
- EC in ML in-situ training?
- Fully integrated on-chip inference

Review: G Wetzstein et al, Nature 588 (2020)

Review: W Bogaerts et al, Nature 586, 207–216 (2020)

S. Bandyopadhyay, Optica 8, 1247 (2021)

R Hamerly et al, Nature Comm. 13 (2023)

S Vadlamani, Science Advances (2023)

Applications

- Computing Across the Internet's Edge: *Tops/sec @mW on edge devices*
- Microlaser-based Coherent Scalable Efficient deep Learning *~ fJ/OP and 25 TeraOP/(mm^2· s)*
- Single-shot NNs: *20,000 MACs / 20 fs: Exascale computing*
- New Tech for High-Speed Spatial Light Modulation
- Multiplicative Analog Frequency Transform on *RF Ph-encoded DNN*

S. Bandyopadhyay, *arXiv:2203.05466* (2022)

A. Sludds et al, Science (2022)

Z. Chen et al, Nature Photonics (2023)

see also Ronald Davis III et al, ArXiv:2207.06883

L Bernstein et al, Science Advances (2023)

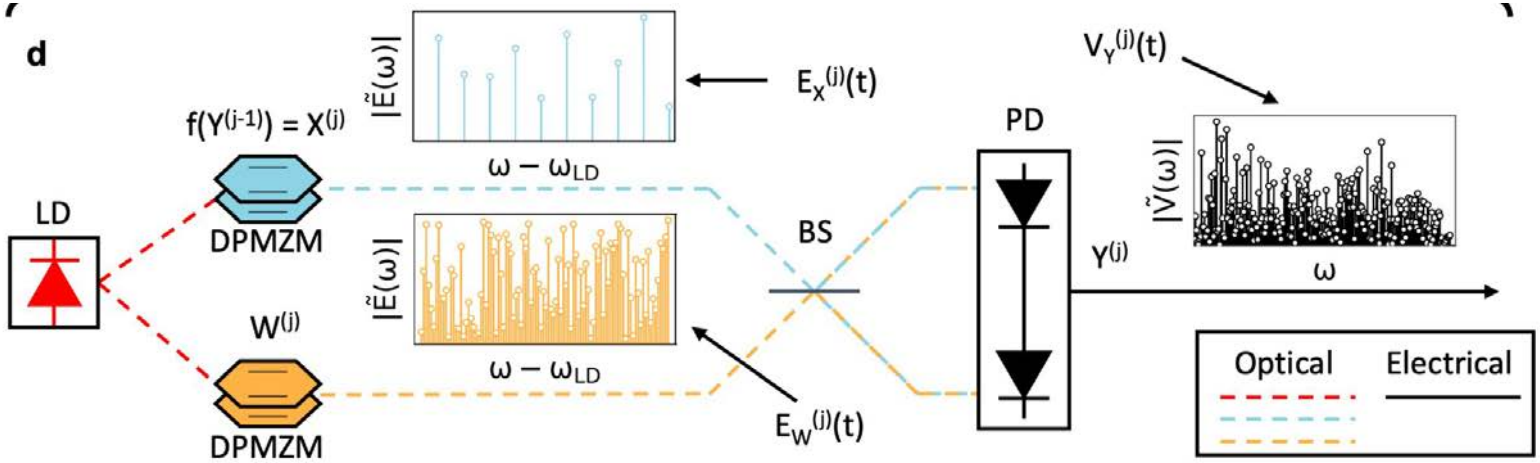
M Dong et al, Nature Photonics (2021); C Panuski et al, Nature Photonics (2023); otehrs in-review (see ArXiv.org)

RF-Photonic Deep Learning Processor with Shannon-Limited Data Movement - Ronald Davis et al, in preparation (2024)

MAFT: A smart optical-RF receiver for ML

All activations X in a layer are encoded into frequency tones

Output Y read out in the output current
Remodulated onto light for next layer

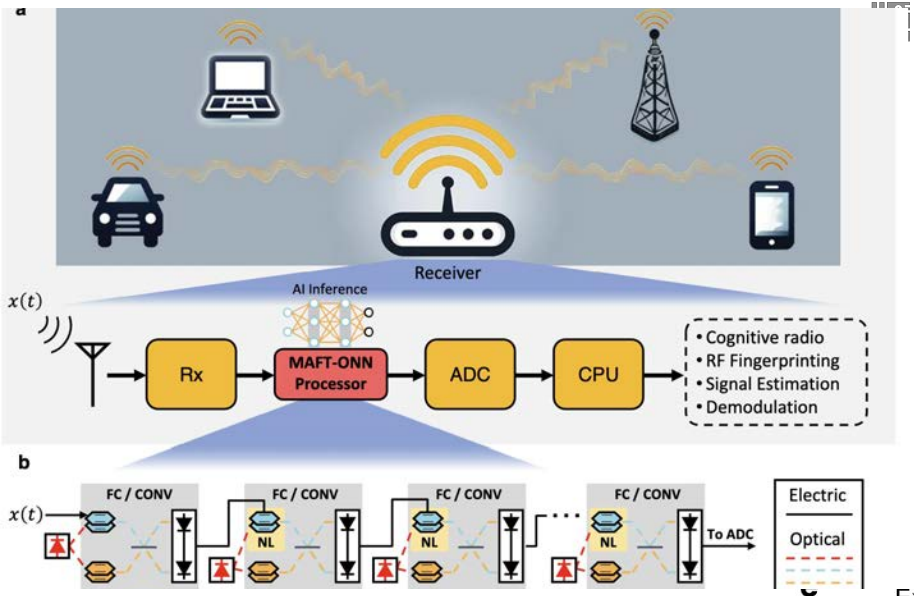


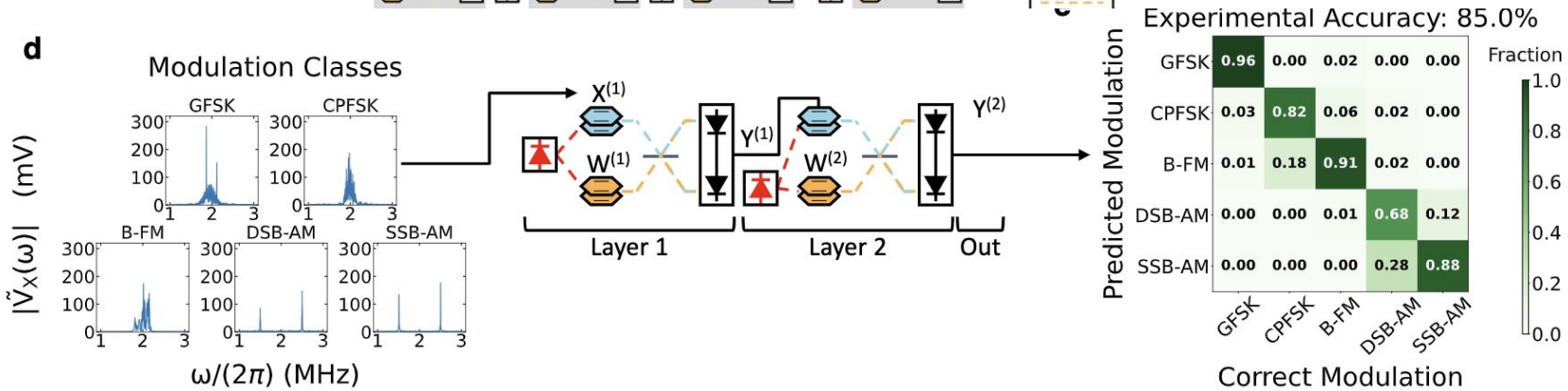
Davis III, R., Chen, Z., Hamerly, R., & Englund, D. (2022). Frequency-encoded deep learning with speed-of-light dominated latency. *arXiv preprint arXiv:2207.06883*.

All layer weights W are encoded into frequency tones

Heterodyne reception gives us one-shot full matrix-vector multiplication for free!

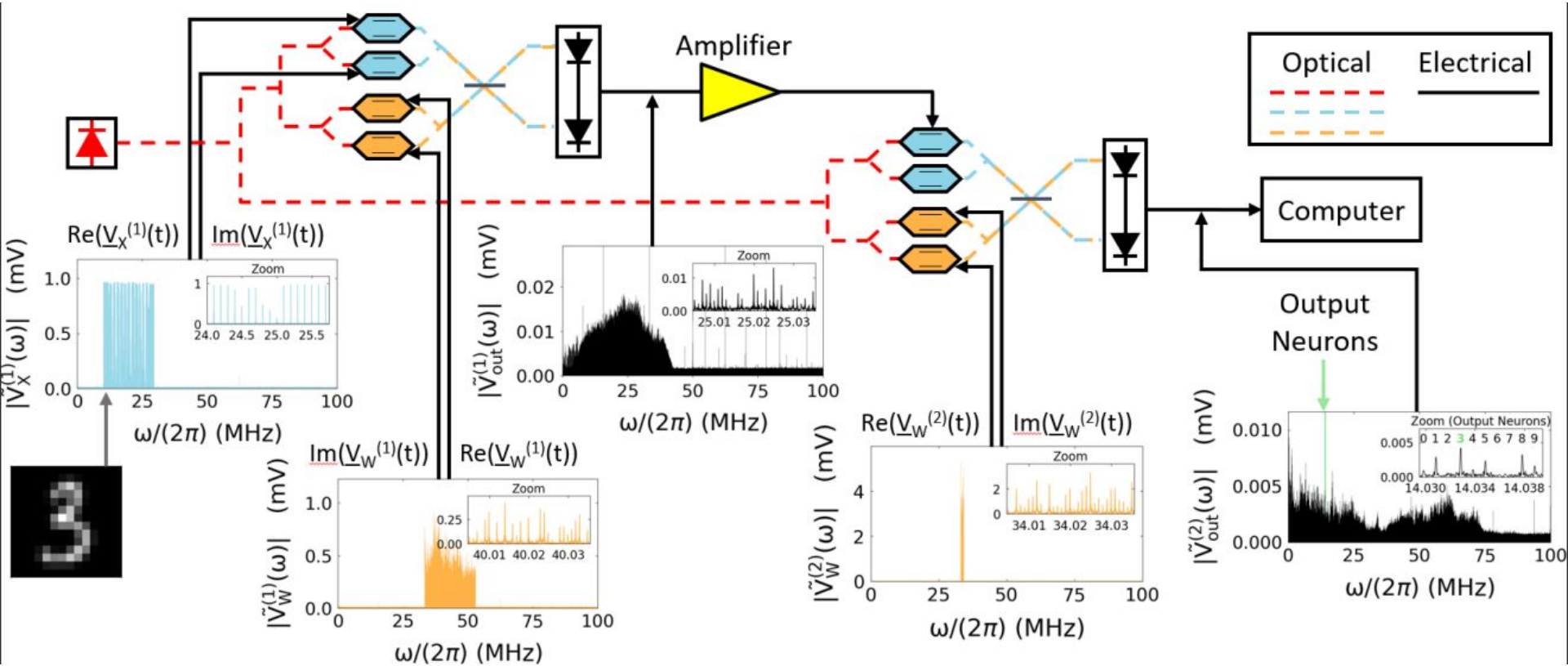
No need to individually add up Multiply ACcumulate operations (MACs)!



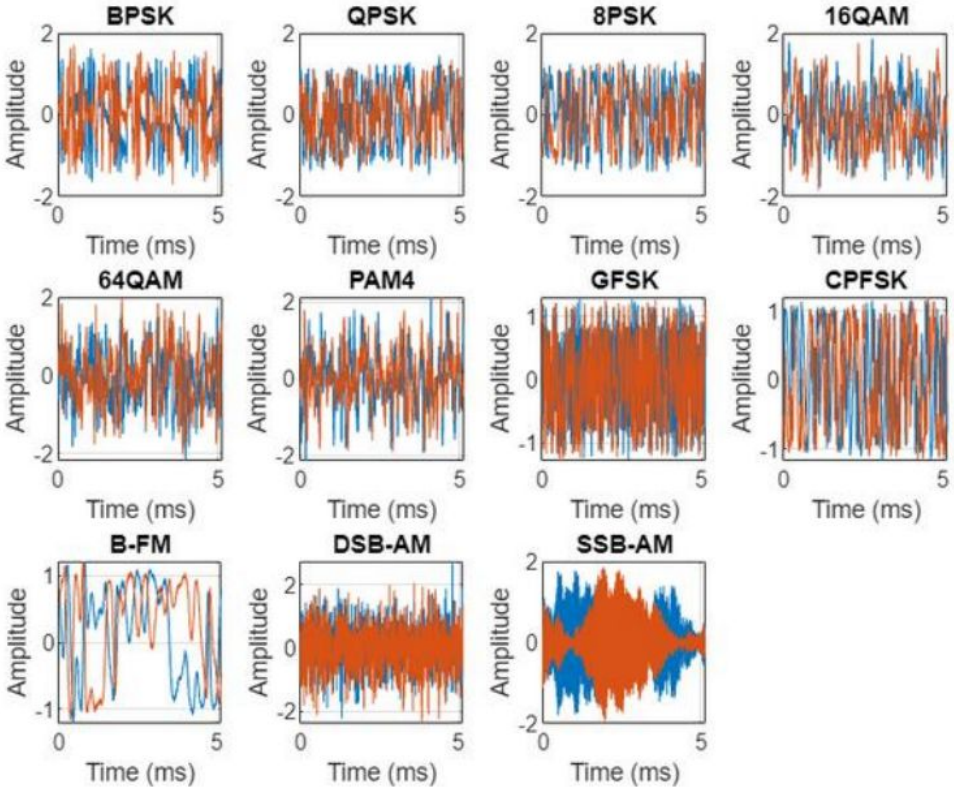


RF-Photonic Deep Learning Processor with Shannon-Limited Data Movement - Ronald Davis et al, in preparation (2024)

Experimental 3-Layer DNN

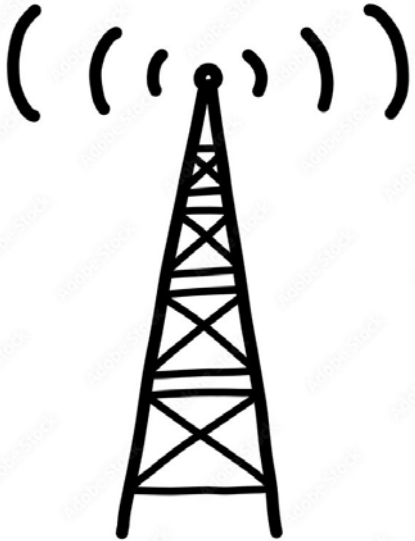


Photonic RFML: Modulation Classification

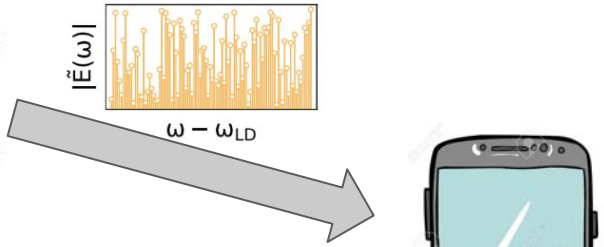


MAFT-ONN Demo

Reprogramming conventional RF devices for ultralow-power ML

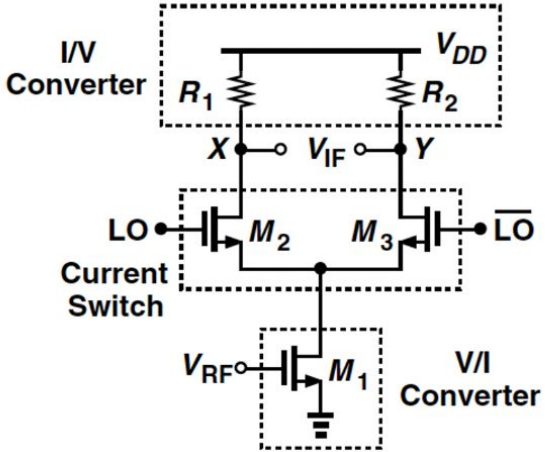


Frequency-encoded weights



Cell phone has local data encoded in frequency domain → this is the “local oscillator” for the mixer

Every cell phone receiver has an RF mixer

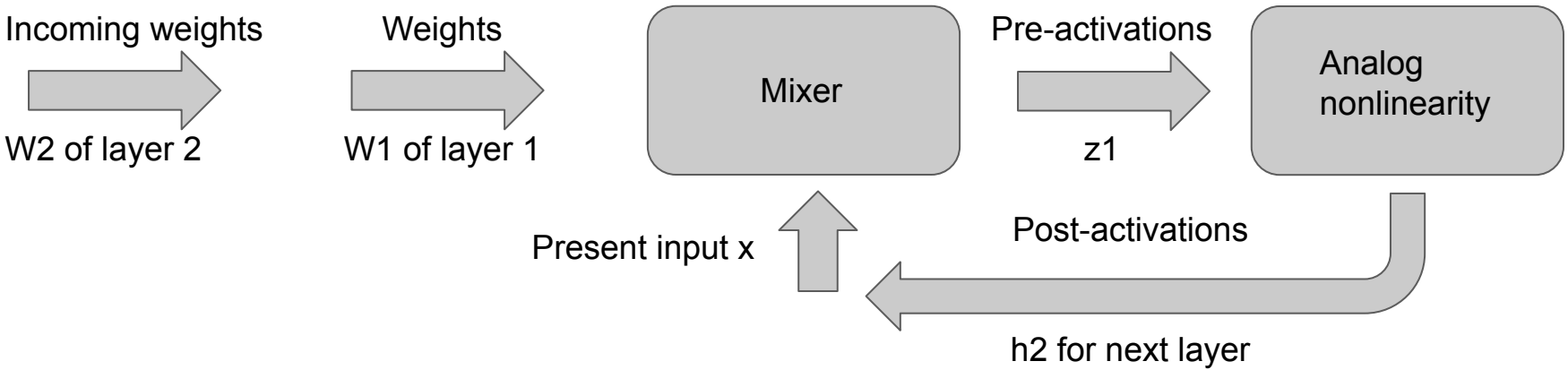


The mixer multiplies the incoming weight signal with the local data signal

*Submitted to MIT Tech Licensing Office

Reprogramming conventional RF devices for ultralow-power ML

The output of the mixer is looped back to the mixer in order to process the next incoming weight layer from the tower.



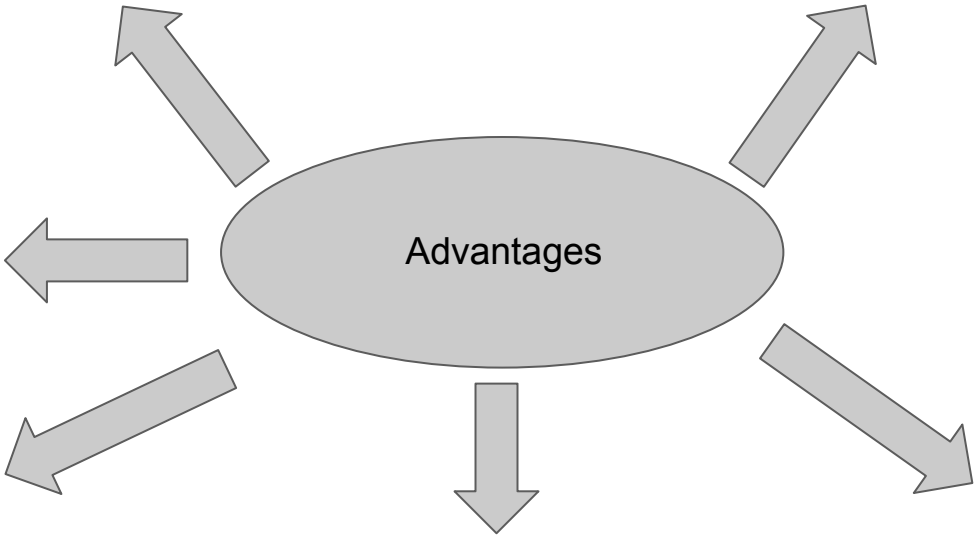
We call this new protocol the Mixed-Signal Extended Range NetCast or the MSER-NetCast scheme

*Submitted to MIT Tech Licensing Office

Machine Intelligence on Wireless Edge Networks (MIWEN*)

Multiplication in time is convolution in frequency
the receiver naturally implements Convolutional Neural Network layers

Standard matrix-vector multiplication recovered by focusing on only certain output frequencies



Simple architecture
No new hardware needed

Extended range from wireless operation

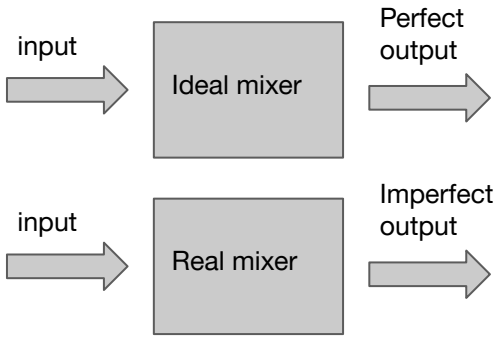
Enables mesh networking of edge devices for crowd/distributed computing/federated learning

Not sharing local inputs with server enables privacy in Healthcare etc.

Machine Intelligence on Wireless Edge Networks (MIWEN*)

Current theoretical efforts

- Deriving and simulating quantum limits of ML on cellphone receivers → performance in photon-starved environments: ultrasmall sensors/antennas or outer space applications
- Performance scaling with temperature, for example, ML accuracy on superconducting Josephson junction mixers
- Performance variation with communication frequency band, 5G, WiFi, 6G



+component noise+thermal noise+quantum fluctuations

Current experimental efforts

- Working with RF groups on a demo of the MSER NetCast scheme both at a small scale and on large RF network testbeds

*Submitted to MIT Tech Licensing Office

Outline

Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing

Review: G Wetzstein et al, Nature 588 (2020)

Review: W Bogaerts et al, Nature 586, 207–216 (2020)

- EC in hardware compute, training: → "error-free optics"

S. Bandyopadhyay, Optica 8, 1247 (2021)

R Hamerly et al, Nature Comm. 13 (2023)

- EC in ML in-situ training?

S Vadlamani, Science Advances (2023)

- Fully integrated on-chip inference

Applications

- Computing Across the Internet's Edge: *Tops/sec @mW on edge devices*

S. Bandyopadhyay, *arXiv:2203.05466* (2022)

- Microlaser-based Coherent Scalable Efficient deep Learning *~ fJ/OP and 25 TeraOP/(mm^2· s)*

A. Sludds et al, Science (2022)

- Single-shot NNs: *20,000 MACs / 20 fs: Exascale computing*

Z. Chen et al, Nature Photonics (2023)

see also Ronald Davis III et al, ArXiv:2207.06883

- New Tech for High-Speed Spatial Light Modulation

L Bernstein et al, Science Advances (2023)

M Dong et al, Nature Photonics (2021); C Panuski et al, Nature Photonics (2023); otehrs in-review (see ArXiv.org)

- Multiplicative Analog Frequency Transform on *RF Ph-encoded DNN*

RF-Photonic Deep Learning Processor with Shannon-Limited Data Movement - Ronald Davis et al, in preparation (2024)

RF-Photonic Deep Learning Processor with Shannon-Limited Data Movement Ronald Davis et al, in preparation (2024)

Outlook

.. with Dr Kurt Jacobs, Dr Dashiel Vittulo, Dr Brian Kirby @ARL

“Programming” the electric field

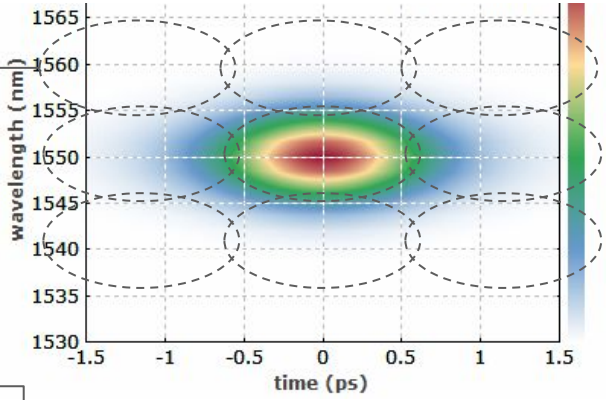
Part 2 - programming the field excitation, $|\psi\rangle = \sum_n \sum_{\boldsymbol{k},\lambda} A(\boldsymbol{k},\lambda,\omega,n)(a^\dagger_{\boldsymbol{k},\lambda})^n |0\rangle$

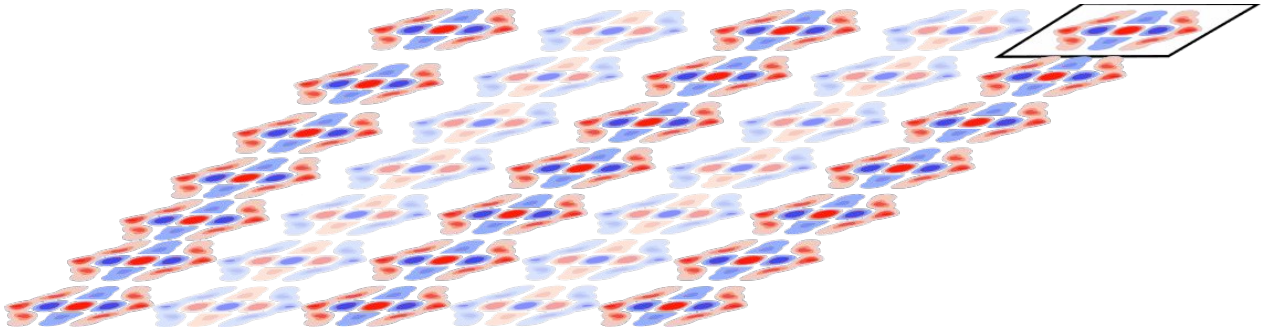
frequency wavelet basis

field operator

$$\boldsymbol{E}(\boldsymbol{r},t) \propto \sum_\lambda \sum_{\boldsymbol{R}} d\omega d^2\boldsymbol{r} A_{\boldsymbol{R},\lambda}(\omega)\hat{a}_{R,\lambda}\boldsymbol{u}_{\boldsymbol{R},\boldsymbol{\lambda}}(\boldsymbol{r}-\boldsymbol{R}) + H.c.$$

position basis

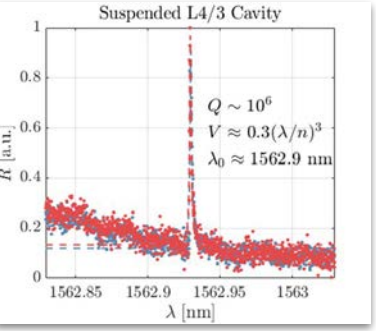


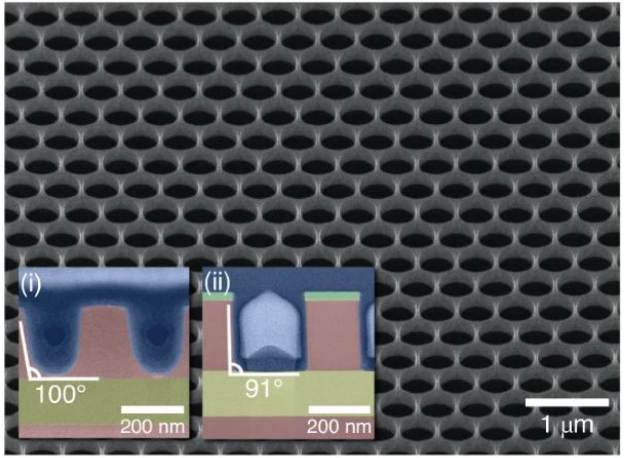


Article | Published: 28 November 2022

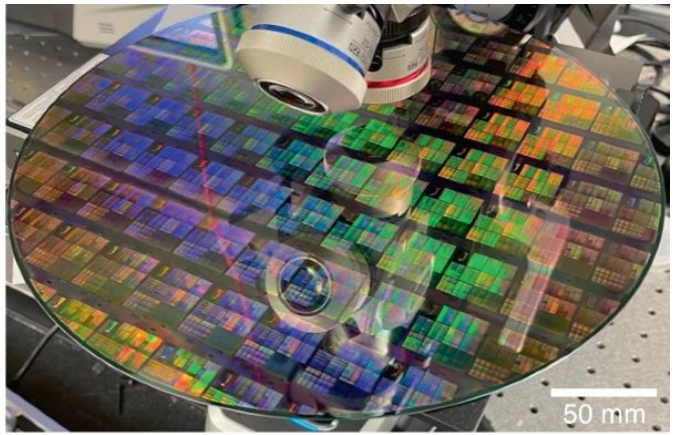
A full degree-of-freedom spatiotemporal light modulator

Christopher L. Panuski ✉, Ian Christen, Momchil Minkov, Cole J. Brabec, Sivan Trajtenberg-Mills, Alexander D. Griffiths, Jonathan J. D. McKendry, Gerald L. Leake, Daniel J. Coleman, Cung Tran, Jeffrey St Louis, John Mucci, Cameron Horvath, Jocelyn N. Westwood-Bachman, Stefan F. Preble, Martin D. Dawson, Michael J. Strain, Michael L. Fanto & Dirk R. Englund ✉

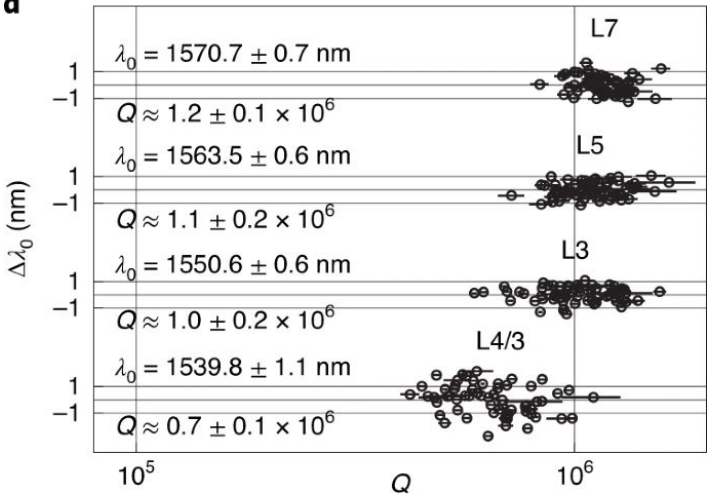




a



d



* Si foundry service at State University of New York Polytechnic

PHYSICAL REVIEW X 10, 041046 (2020)

Featured in Physics

Demonstrate $\delta T_{min} < \mu K/(Hz)^{1/2}$ in nanoscale cavity

⇒ further analysis shows strong promise for qu-limited bolometry focal plane arrays (U.S. Application No. 63/141,646)

Fundamental Thermal Noise Limits for Optical Microcavities

Christopher Panuski,[1,*] Dirk Englund,[1] and Ryan Hamerly[1,2,†]
[1]Research Laboratory of Electronics, MIT, Cambridge, Massachusetts 02139, USA

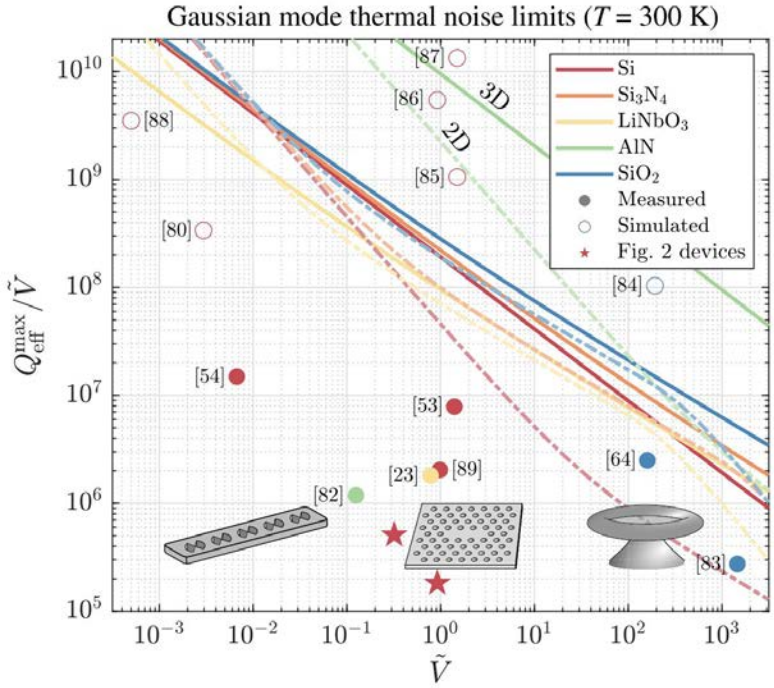
Relevant upshot: "Theoretical + experimental analysis of thermorefractive noise in high-quality-factor (Q), small-mode-volume (V) optical microcavities. Analogous to well-studied stability limits imposed by Brownian motion in macroscopic Fabry-Perot resonators, we show that microcavity thermorefractive noise gives rise to a mode-volume-dependent maximum effective quality factor.. [..]" -- and we offer a solution to go beyond

$$\langle \tilde{a}(t)\tilde{a}^*(t+\tau)\rangle \sim e^{-(\Gamma+\delta\omega_{\mathrm{rms}}^2/\Gamma_T)\tau}$$

where the thermal mode volume

$$V_T = \frac{[\int \epsilon(\vec{r})|\vec{E}(\vec{r})|^2 d^3\vec{r}]^2}{\int \epsilon(\vec{r})^2|\vec{E}(\vec{r})|^4 d^3\vec{r}}$$

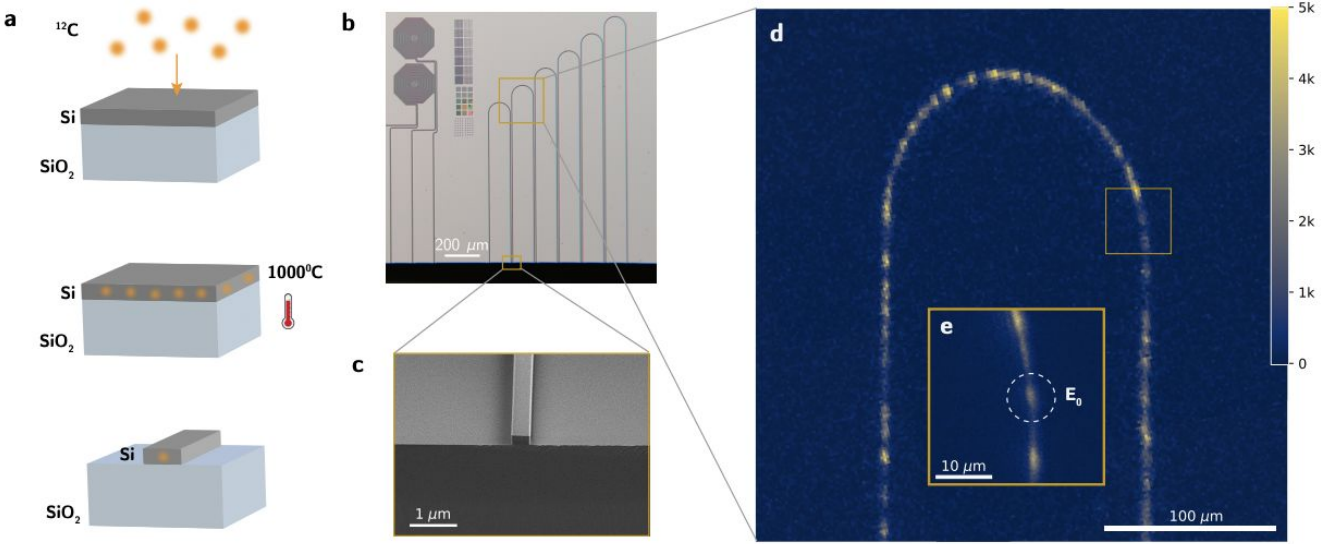
is the common Kerr nonlinear mode volume

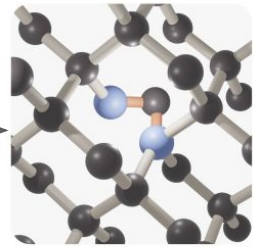


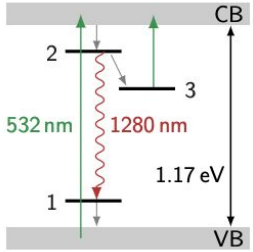
→ high-speed bolometry/C Panuski, J Goldstein, et al :supported by Army Research Lab ISN

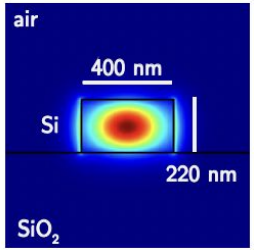
Individually Addressable Artificial Atoms in Silicon Photonics

carbon-related ‘G-center’



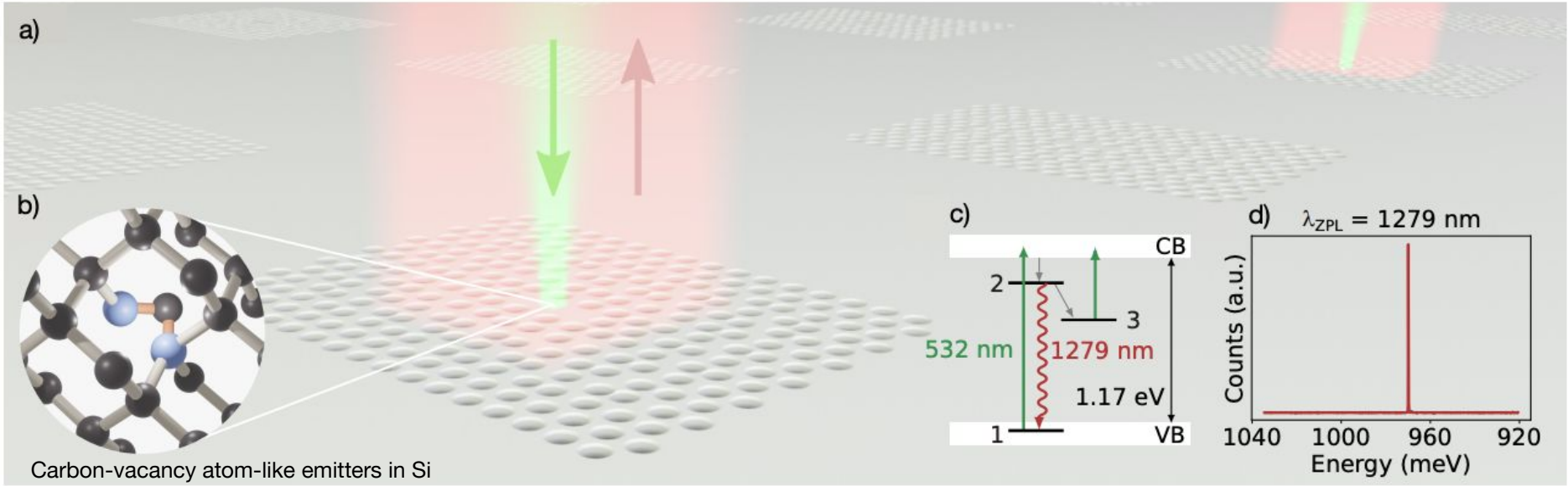






M. Prabhu, C. Errando Herranz, et al *Nature Communications* (2023).

SLMs controlled by individual atomic memories

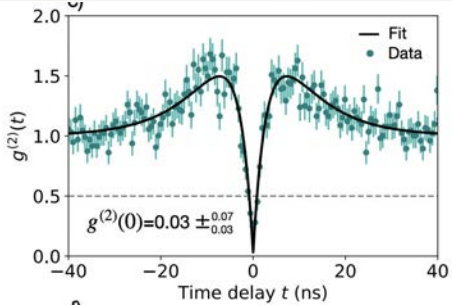


Carbon-vacancy atom-like emitters in Si

But wait, haven't we seen this playbook before?

Purcell SE enhancement in QD-PhC: DE et al (Vuckovic), PRL **95** (2005) ⇒ single QD-controlled reflectivity (DE,Faraon, Fushman, ..Vuckovic Nature 450, 2007), etc.. → Many other ss cavity-QED systems, such as SiV-cavity (Lukin, Loncar, Englund, Vuckovic, etc)

But all-Si platform opens up huge possibilities .

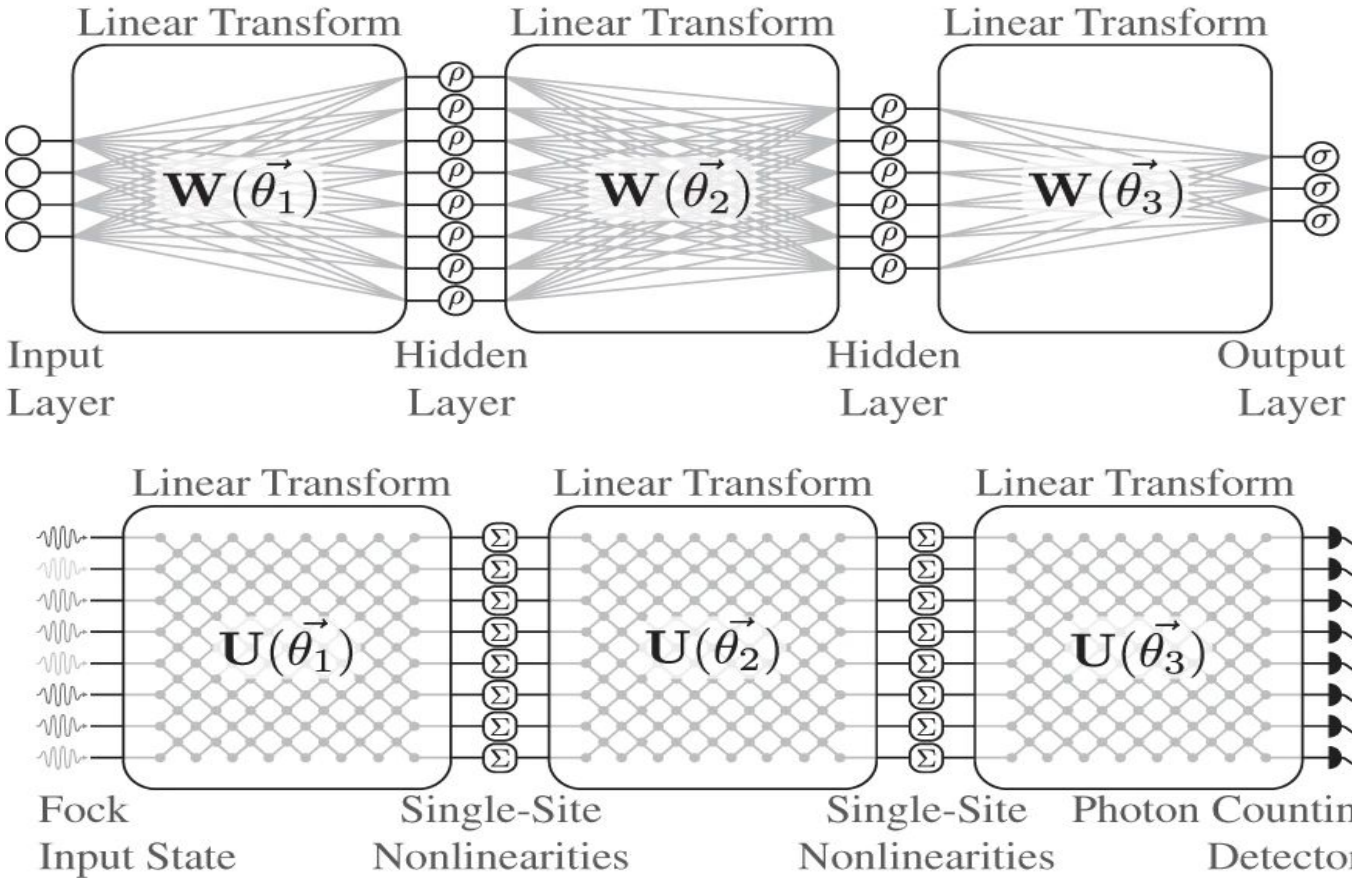


*Valeria Saggio, Carlos Errando-Herranz, et al arXiv:2302.10230v1 (2023)

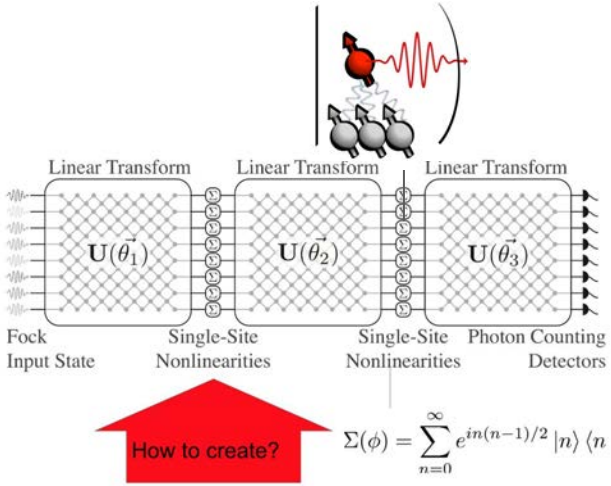
What could a DNN do with a *quantum* nonlinearity?

information flow:

physics implementation:



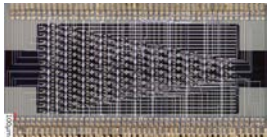
A Quantum Optical Neural Network (QONN)



*

Supervised training[1] →

- Quantum optical state compression
- Reinforcement learning
- Black-box quantum simulation
- One-way quantum repeaters



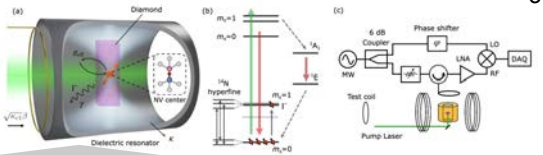
Nonlinearities: atoms, artificial atoms, QDs, molecules, or bulk optical NL[2,3]

1. G. Steinbrecher et al, NPJ Quantum Information Processing **5** (2019)
2 M Heuck, K. Jacobs (ARL), DE, PRL **124** (2020); 3. S. Krastanov et al, N COMM **12** (2021)

Distributed Deep Learning + quantum (-limited) sensing & networking

MIT, MIT LL, ARL, ADI, Honda RI

δB_k <pT/Hz$^{1/2}$



weights W_{nmk}

output $f_W(a_{nmk})$

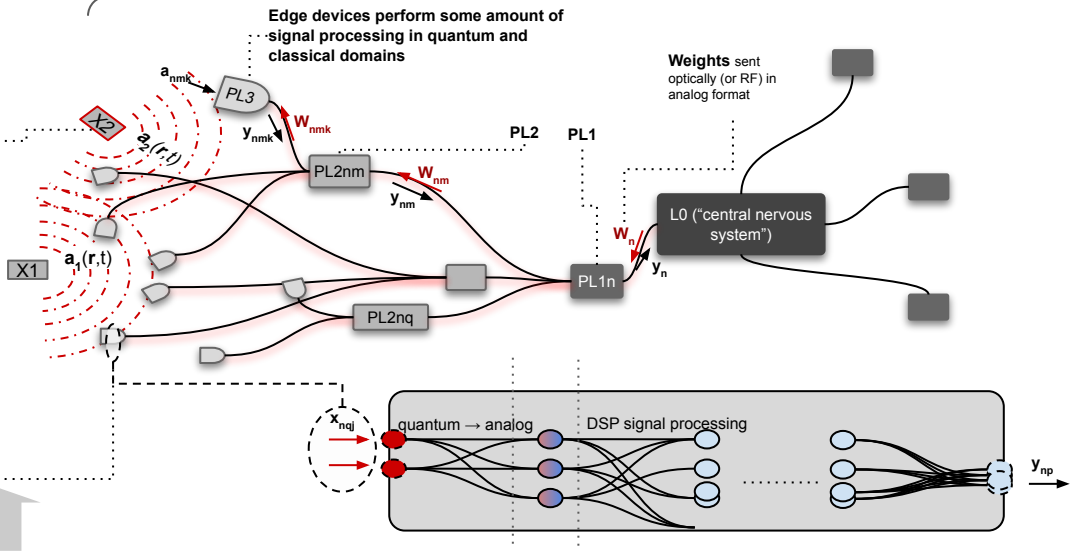
EM signal (e.g., microwave transmission through spin sensor ro reflected radar from object X_2=UAV)

Some open questions:

*- Need to understand interplay of **ML "software"** with microsystems ("brain" + form)*

- Narrowing focus: Co-design system S: quantum hardware $\Theta_{hardware}$ & machine intelligence $\Theta_{algorithm}$? Where to place quantum-analog-digital boundaries?

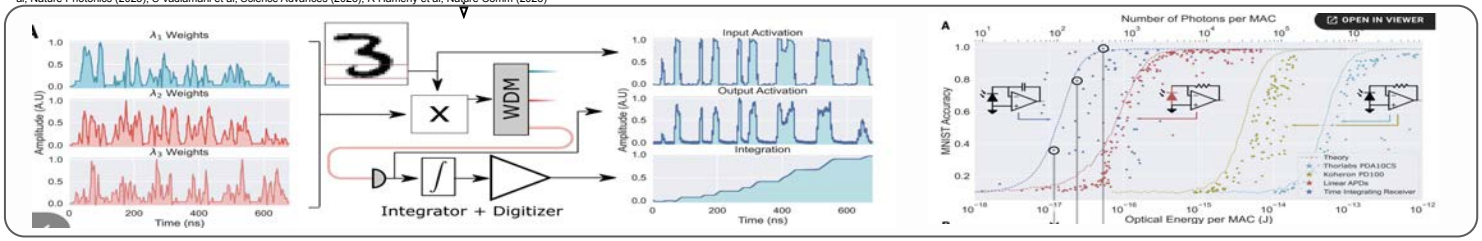
System S ∈{autonomous vehicle, UAV network, quantum computer, radar installation, ..}



weights W_{nmk}

output $f_W(a_{nmk})$

Squeezed light resource can beat standard quantum limit. Quantum-limited programmable matrix algebra (vector products, convolutions, etc). Specifically, computing a_N . b_N is O(1) operations vs O(N) operations in DSP. A Sludds et al, Science (2022); L Bernstein et al, Science Advances (2023); Z. Chen et al, Nature Photonics (2023); S Vadlamani et al, Science Advances (2023); R Hamerly et al, Nature Comm (2023)



Positions available in theory & experiment: https://www.rle.mit.edu/qp/

Summary & Closing Remarks

Intro

- A Renaissance in Physics of Computing ?

Hardware

- In-physics processing

Review: G Wetzstein et al, Nature 588 (2020)

Review: W Bogaerts et al, Nature 586, 207–216 (2020)

- EC in hardware compute, training: → "error-free optics"

S. Bandyopadhyay, Optica 8, 1247 (2021)

R Hamerly et al, Nature Comm. 13 (2023)

- EC in ML in-situ training?

S Vadlamani, Science Advances (2023)

- Fully integrated on-chip inference

Applications

- Computing Across the Internet's Edge: *Tops/sec @mW on edge devices*

S. Bandyopadhyay, *arXiv:2203.05466* (2022)

- Microlaser-based Coherent Scalable Efficient deep Learning *~ fJ/OP and 25 TeraOP/(mm^2· s)*

A. Sludds et al, Science (2022)

- Single-shot NNs: *20,000 MACs / 20 fs: Exascale computing*

Z. Chen et al, Nature Photonics (2023)

see also Ronald Davis III et al, ArXiv:2207.06883

- New Tech for High-Speed Spatial Light Modulation

L Bernstein et al, Science Advances (2023)

M Dong et al, Nature Photonics (2021); C Panuski et al, Nature Photonics (2023); otehrs in-review (see ArXiv.org)

- Multiplicative Analog Frequency Transform on *RF Ph-encoded DNN*

RF-Photonic Deep Learning Processor with Shannon-Limited Data Movement - Ronald Davis et al, in preparation (2024)

Outlook

.. with Dr Kurt Jacobs, Dr Dashiel Vittulo, Dr Brian Kirby @ARL

ıre Pr The Washington Post
Democracy Dies in Darkness

Amid explosive demand, America is running out of power

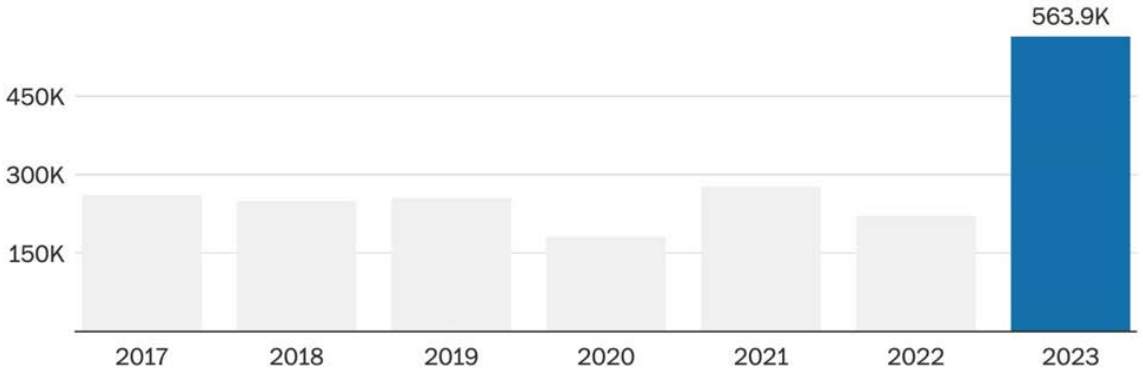
AI and the boom in clean-tech manufacturing are pushing America's power grid to the brink. Utilities can't keep up.

By Evan Halper

March 7, 2024 at 6:05 a.m. EST

Projected new energy demand in North America doubles

9-year growth forecast of demand for new electricity, in gigawatt hours



Data covers U.S., Canada and part of Baja California, Mexico.

Source: North American Electric Reliability Corp. Long Term Reliability Assessment

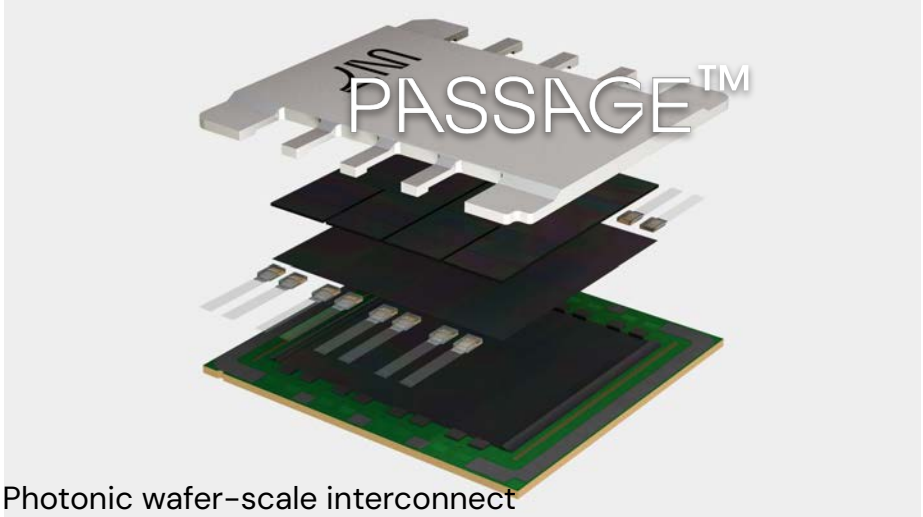
Addressing the ML bottleneck: LightMatter

PASSAGE™

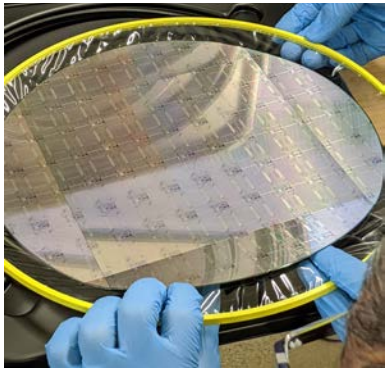
A wafer-scale, programmable photonic interconnect that enables arrays of heterogeneous chips to communicate.

IDIOM®

Standard deep learning frameworks and model exchange formats, providing the transformations and tools required.



Photonic wafer-scale interconnect



PERFORMANCE

40X
Dramatic interconnect density improvement. 40 waveguides in the space of one optical fiber.

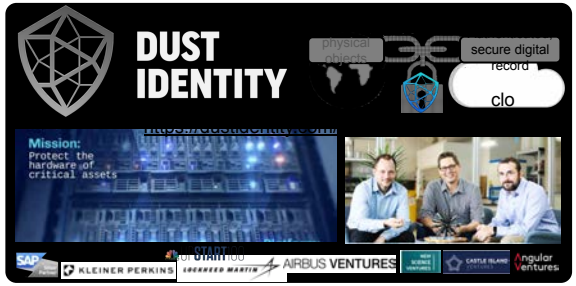
800+ TBPS
Input/output bandwidth from each chiplet site for full reticle. And up to 250+ Tbps per chiplet site edge.

100X
More bandwidth than existing chip-to-chip interconnect solutions.

<2NS
Chiplet to chiplet latency, single hop connectivity between every site.

*disclaimer: Englund is scientific advisor to LM

Basic research → **Technology transition** → Industry, gov't partnerships → **Startups**



Quantum Network Technologies, Inc

Quantum communications, interconnects, memories

Laura Andre, Ruo Li
[with F Klein]

Englund - Soljacic - Koppens - Aspuru-Guzik - [redacted] - [redacted] .
Launching summer 2024
Now hiring

Summary

Key point #1: **Try to use data fan-out** for “free” data-movement

Key point #2: **In-physics computing** (allows inner vector products with O(1) operations)

Key point #3: **Coherent error cancellation** possible in analog optical systems

Key point #4: **In-situ training**

- single-shot $\mathbf{f}_{DNN}(\mathbf{x})$ evaluation → more general than back-prop
- transferable learning

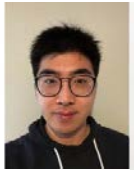
Acknowledgements (ML portion)

MIT Quantum Photonics Group: Dr. Ryan Hamerly (NTT, MIT)

PhD: Saumil Bandyopadhyay, Nick Harris (--> LightMatter), Darius Bunandar (--> LightMatter), Mihika Prabhu, Chris Panuski, Liane Bernstein, Alex Sludds, Ronald Davis

Postdocs: Stefan Krastanov, Sri Krishna Vadlamani, Zaijun Chen (--> USC)


Saumil Bandyopadhyay


Prof. Zaijun Chen


Dr. Nicholas Harris


Dr. Zhizhen Zhong


Liane Bernstein


Dr. Stefan Krastanov


Dr. Carlos Herranz


Sivan Trajtenberg-Mills


Hugo Larocque


CPT Dr Chris Panuski (--> U.S. Navy)

Collaborators on this work:

MIT: Manya Ghobadi, Joel Emer, Vivienne Sze

MITRE Corp: Gerry Gilbert, Mark Dong, Gen Clark
ARL: Dr Dashiel Vittulo, Dr Kurt Jacobs

Sandia NL: M Eichenfield
AFRL: Dr Mike Fanto, et al
Prof Tingjun Chen (Duke), Saikat Guha (UMD), M Eichenfield (U. Arizona)


Alexander Sludds


Ronald Davis III

Dr Sri Vadlamani


Prof. Manya Ghobadi


Dr. Ryan Hamerly

Funding

AFOSR AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
NSF
NTT
ANALOG DEVICES
SCIENCE TO SHAPE THE FUTURE OF THE ARMY · ARMY RESEARCH OFFICE · 1951
MITRE
HRI Honda Research Institute US
tsmc
DARPA
MTL MICROSYSTEMS TECHNOLOGY LABORATORIES