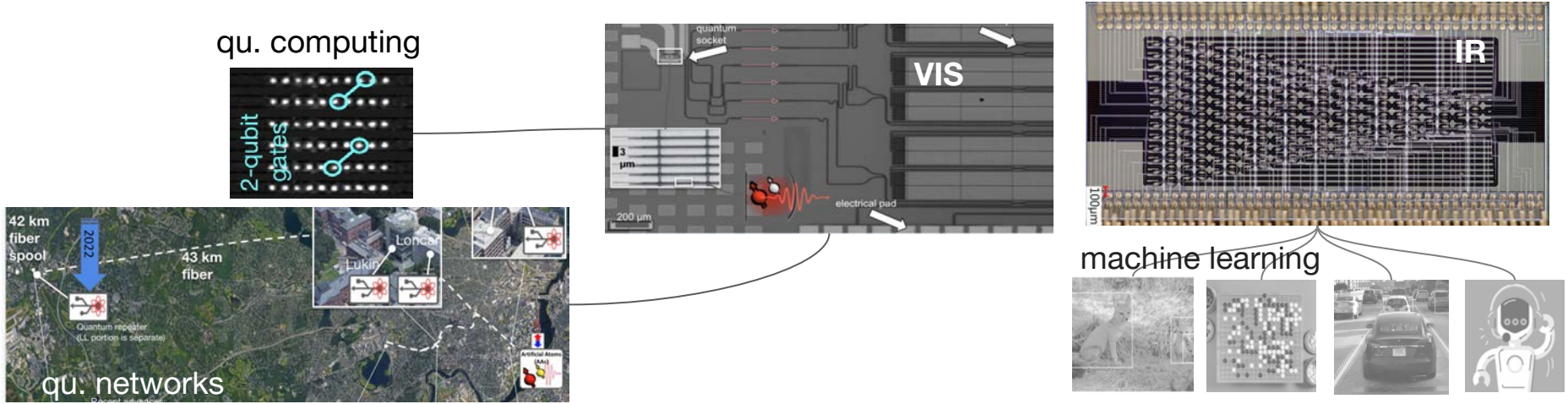
Programming optics for quantum & neuromorphic computing

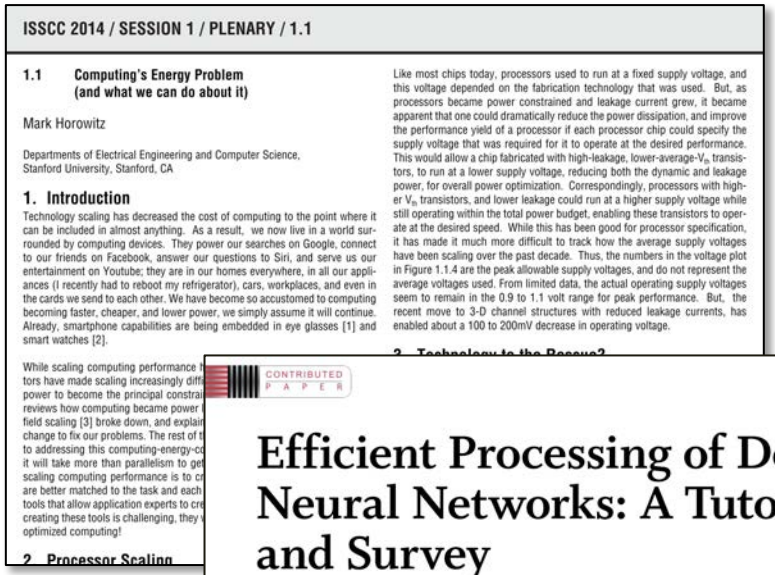
AI Hardware Program Fall Research Update

Dirk Englund*

* ∈ {RLE, MTL, EECS}@MIT Quantum Photonics Laboratory @ MIT
& Collaborators ∈ {MIT, MIT-LL, Harvard, Brookhaven National Laboratory, Sandia National Laboratory, U-Arizona, Stanford, BBN Technologies, MITRE Corp, U. Vienna, Air Force Research Laboratory@Rome, Army Research Laboratory, Oak Ridge NL, U-Maryland,QuEra Computing, TU Berlin, QuEra}



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.

While scaling computing performance f... tors have made scaling increasingly diffi... power to become the principal constrai... reviews how computing became power ... field scaling [3] broke down, and explai... change to fix our problems. The rest of t... to addressing this computing-energy-co... it will take more than parallelism to ge... scaling computing performance is to cr... are better matched to the task and each ... tools that allow application experts to cre... creating these tools is challenging, they ... optimized computing!

2. Processor Scaling

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 deployment of DNNs in AI systems. This article aims to provide... 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"

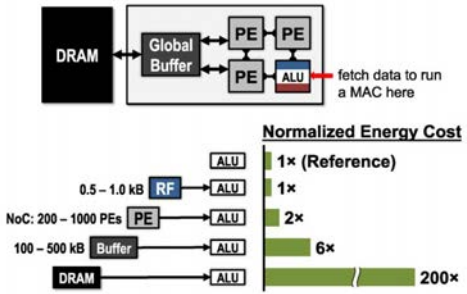


Fig. 22. ***Memory hierarchy and data movement energy [82].***

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

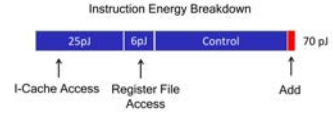
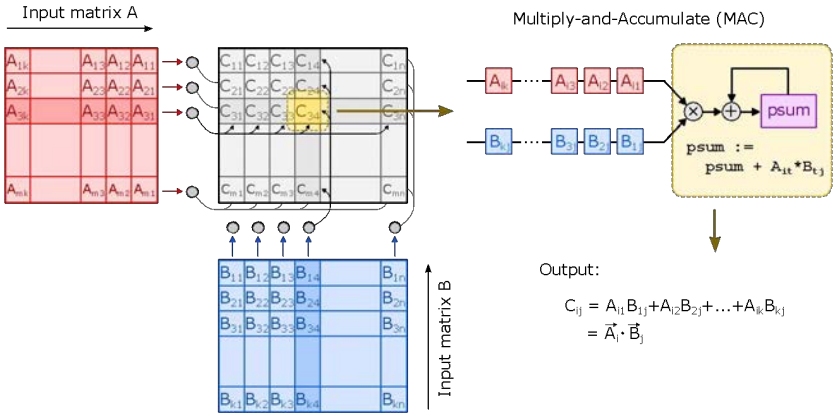


Figure 1.1.9: Rough energy costs for various operations in 45nm 0.9V.

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

slide: Ryan Hamerly

Systolic Array architecture is optimized for GEMM



Two key limiters to the TPU:

- Power ⟶ Heat dissipation
- Interconnects / Data Movement

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

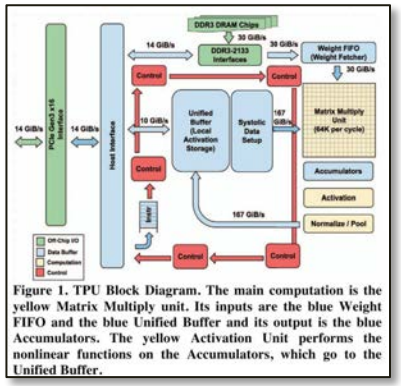


Figure 1. TPU Block Diagram. The main computation is the yellow Matrix Multiply unit. Its inputs are the blue Weight FIFO and the blue Unified Buffer and its output is the blue Accumulators. The yellow Activation Unit performs the nonlinear functions on the Accumulators, which go to the Unified Buffer.

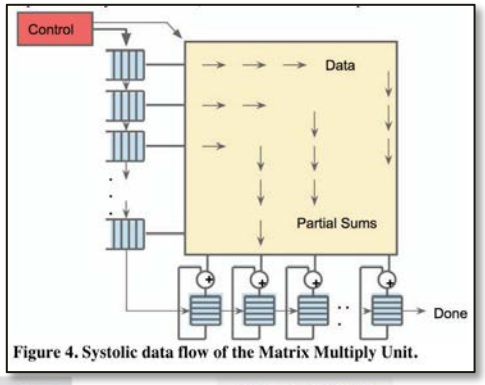
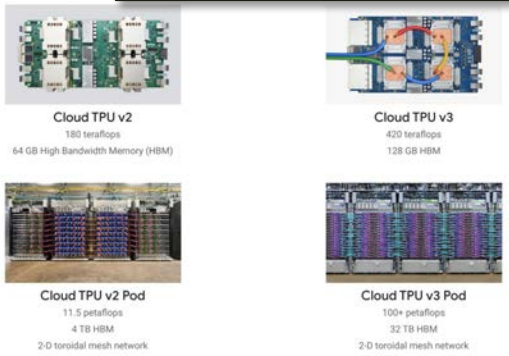


Figure 4. Systolic data flow of the Matrix Multiply Unit.

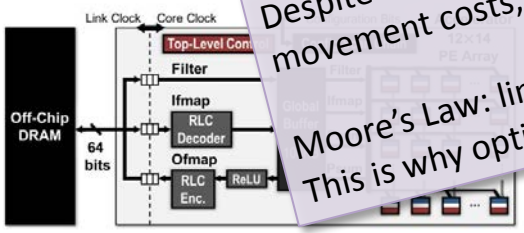
Google TPU cloud offerings:



How does this matter for actual DNNs?

Start with a DNN accelerator design (such as Eyeriss)

- Not just a systolic array!
- Has a balanced hierarchy of memory to handle many tasks

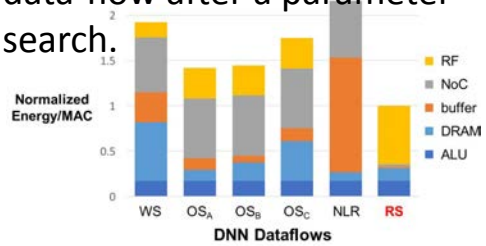


Despite all the optimizations, chips are dominated by data movement costs, not raw processing!

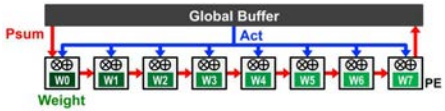
Moore's Law: limited gains for data movement
This is why optics is interesting for future DNN hardware

Find the optimal dataflow

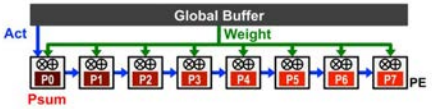
- Trying to minimize data movement. But the data must
- Schemes: "weight-stationary" (systolic), "output-stationary"
- A compiler picks the best data-flow after a parameter search.

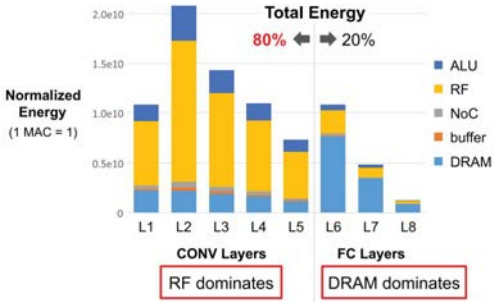


(weight-stationary)



(output-stationary)



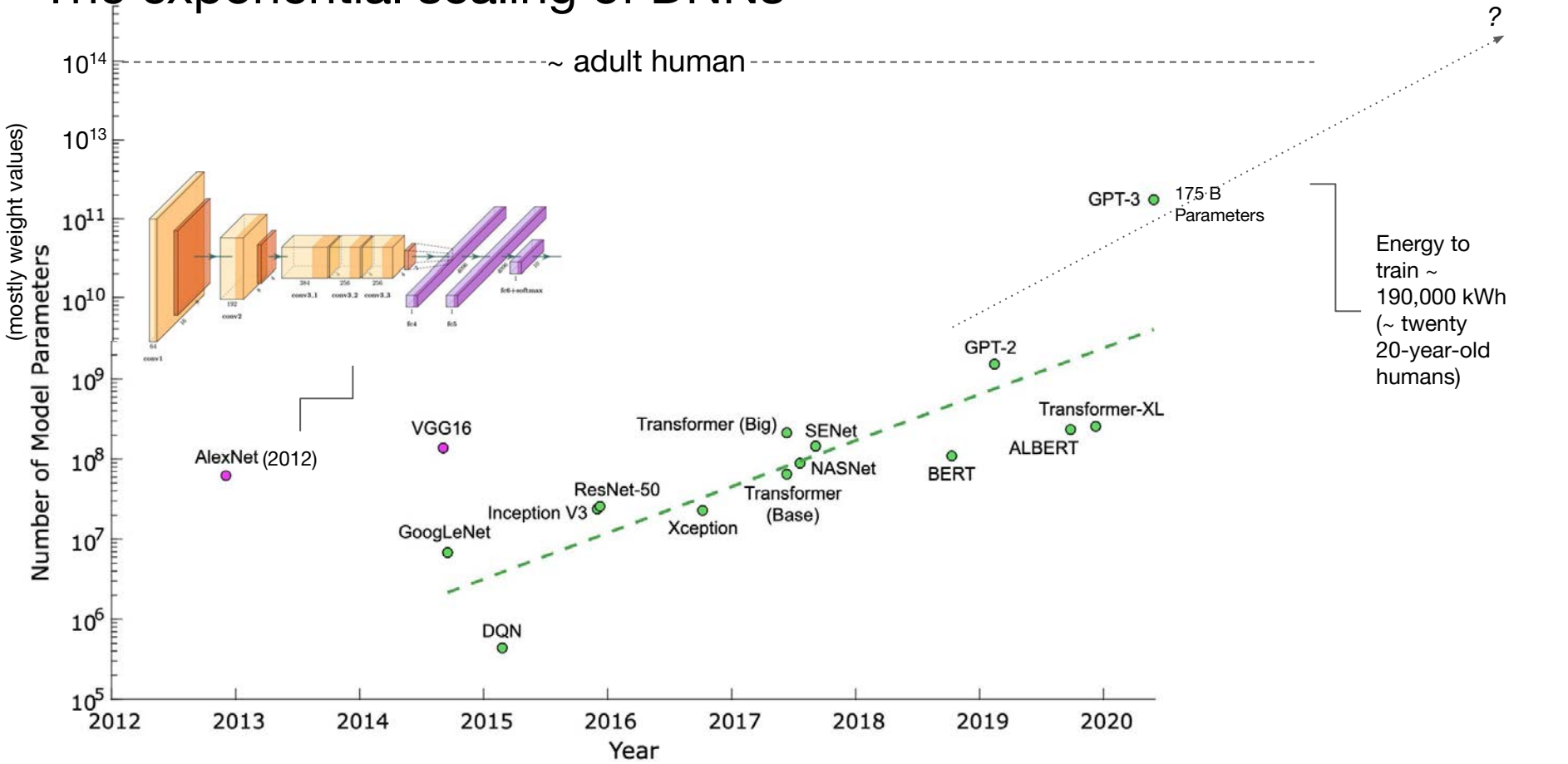


V. Sze, Proc. IEEE 105(12), 2295 (2017)

slide: Ryan Hamerly

Motivation:

The exponential scaling of DNNs



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

Outline

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

Intro

- A Renaissance in Physics of Computing ?

Hardware

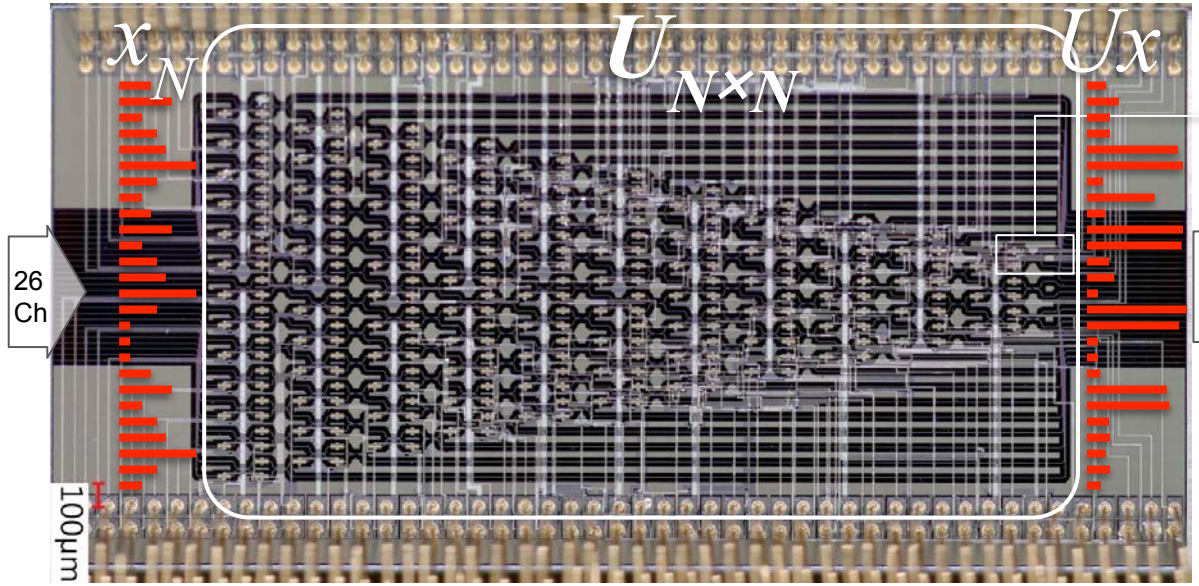
- Programmable photonic integrated circuits (PICs)

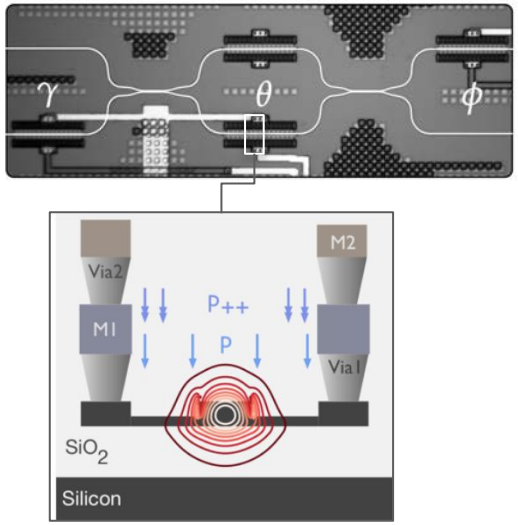
Review: G Wetzstein et al, Nature 588 (2020)
Review: W Bogaerts et al, *Nature 586, 207–216*

Applications

Example:

Matrix algebra in programmable photonic integrated circuits (PICs)





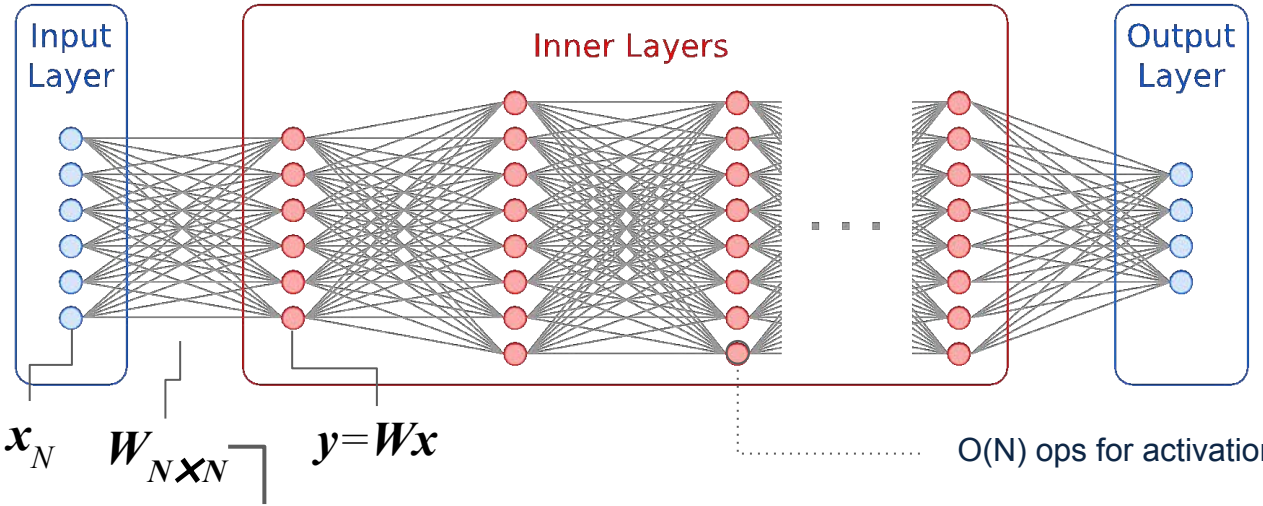
Programmable optical matrix transformation: 176 phase shift degrees of freedom (8 bit)

Thermal phase shifters burn ~ 10 mW/each, but we have piezoelectric alternative (M Dong et al - Nature Photonics).. and nonvolatile phase shifters as by Pernice group are super interesting if we could program precisely

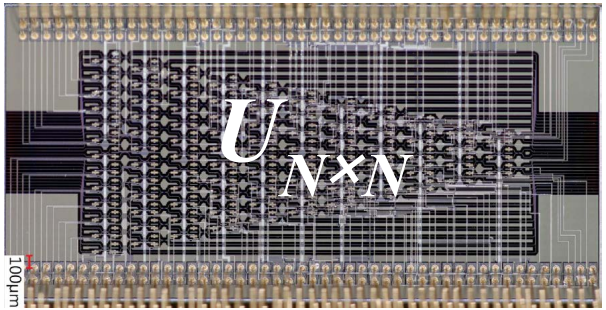
- Y. Shen*, N. C. Harris*, et al [with M Soljacic, MIT], Nature Photon **11** (2017). * equal authors
- See also D.A.B.M Miller, “Sorting out Light” , Science **347** (2017)
- Reviews:
  - N Harris et al, Optica **5** (2018)
  - W Bogaerts, D Pérez, J Capmany, D A. B. Miller, J Poon, D Englund, F Morichetti & A Melloni, **Nature 586** (2020)
  - G Wetzstein, A Ozcan, S Gigan , S Fan, D Englund, M Soljačić, C Denz, D A. B. Miller, D Psaltis, **Nature 588** (2020)

OPSIS Foundry
Collaborators: Michael Hochberg, Michael Fanto, Paul Alsing (AFRL), Stefan Preble (RIT), Philip Walther (U. Vienna)

“Weight-stationary architecture”: flying output **y**=(stationary **W**).(flying **x**)



→ MIT spinouts



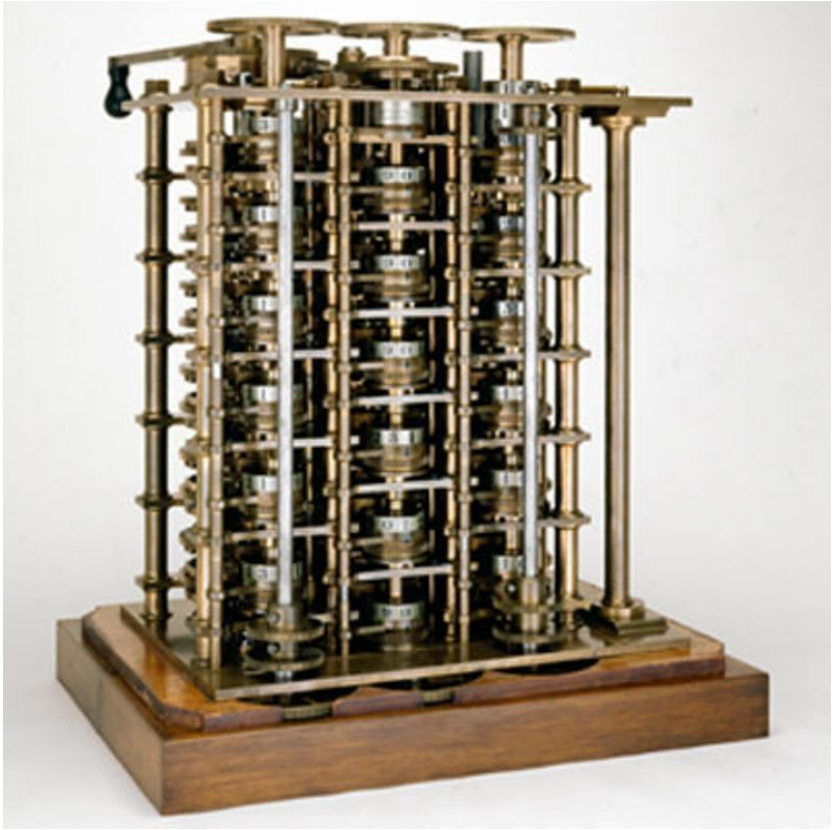
$U_{N\times N}\,x_N$ vector-matrix multiplication:

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

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

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

- Programmable photonic integrated circuits (PICs)

  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)
  Hamerly et al., arXiv:2106.03249 (2021)
  R Hamerly, S Bandyopadhyay, DE, arXiv: 2109.05367 (2021)

Applications

Error correction techniques in Mach Zehnder Meshes

Numerical Optimization

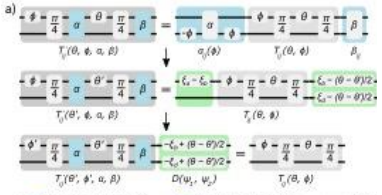
Nonlinear optimizer: eg minimize norm $|U - U_0|$

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

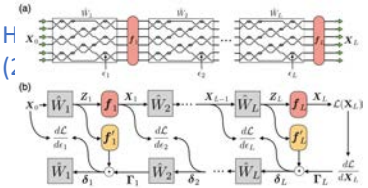
Given errors, find (θ,φ) that fix them per-MZI.



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

In-situ Training

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

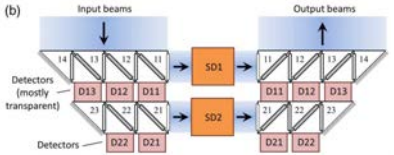


(+) Arbitrary meshes
(–) Slow, per-device training
(+) No pre-calibration
(–) Additional $O(N^2)$ detectors

Progressive Methods

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

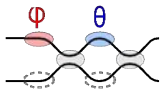
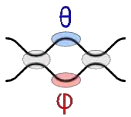
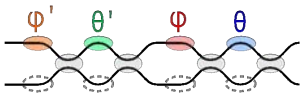
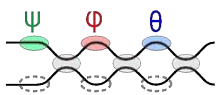
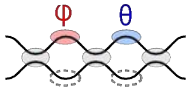
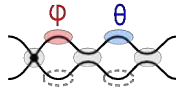
Miller et al., OE 21, 6360 (2013)
Miller et al., Phot Res 1, 1 (2013)



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

Question: Can one construct “perfect” photonics from imperfect components? And can it be done efficiently?

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)
Hamerly et al., arXiv:2106.03249 (2021)
Ryan Hamerly, Saumil Bandyopadhyay, DE, arXiv: 2109.05367 (2021)

* 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

- Programmable photonic integrated circuits (PICs)
- 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)
Hamerly et al., arXiv:2106.03249 (2021)
R Hamerly, S Bandyopadhyay, DE, arXiv: 2109.05367 (2021)

Applications

- Fully integrated on-chip inference

arXiv:2203.05466 (2022)

See Saumil's talk coming up

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*

Outline

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

Intro

- A Renaissance in Physics of Computing ?

Hardware

- Programmable photonic integrated circuits (PICs)
- 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)
Hamerly et al., arXiv:2106.03249 (2021)
R Hamerly, S Bandyopadhyay, DE, arXiv: 2109.05367 (2021)

Applications

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

arXiv:2203.05466 (2022)

ArXiv:cs.ET cs.LG 2205.09103
(2022); to appear in Science

Hey Siri

See Alex Sludd's talk coming up

Outline

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

Intro

- A Renaissance in Physics of Computing ?

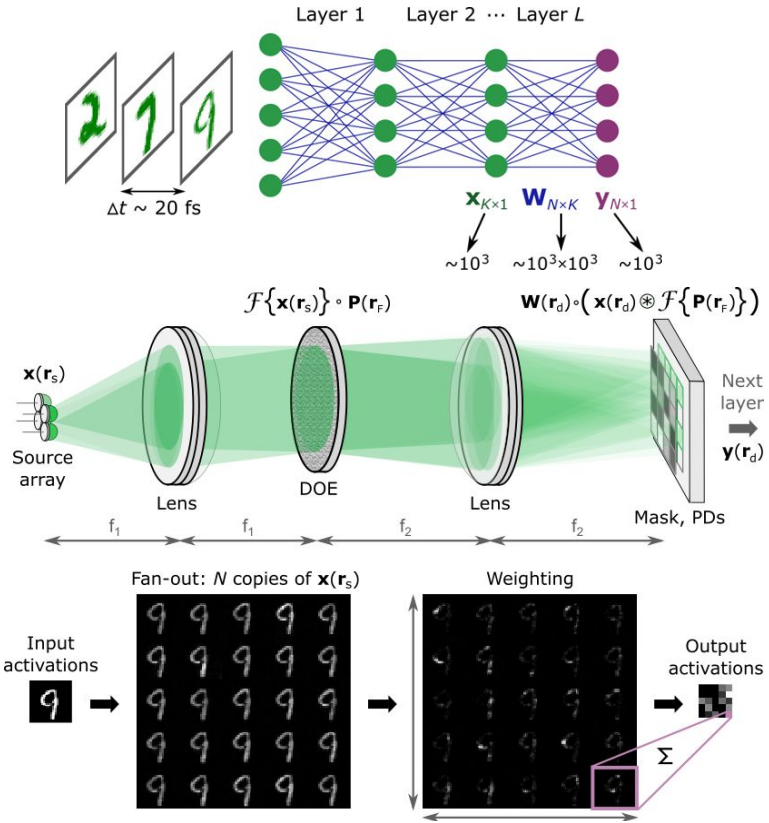
Hardware

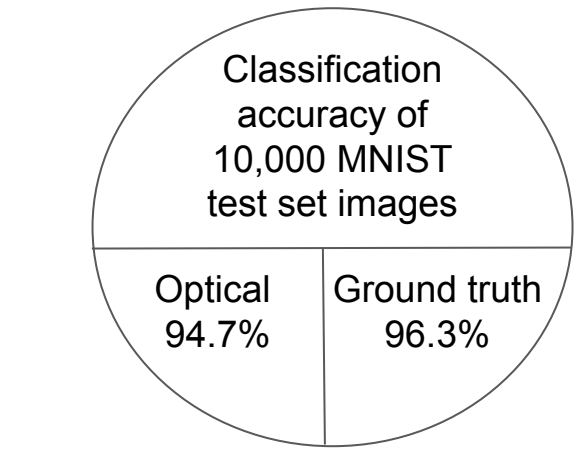
- Programmable photonic integrated circuits (PICs)
  Review: G Wetzstein et al, Nature 588 (2020)
  Review: W Bogaerts et al, *Nature 586, 207–216* (2020)
- EC in hardware compute, training: → "error-free optics"
  S. Bandyopadhyay, Optica 8, 1247 (2021)
  Hamerly et al., arXiv:2106.03249 (2021)
  R Hamerly, S Bandyopadhyay, DE, arXiv: 2109.05367 (2021)
- EC in ML in-situ training?

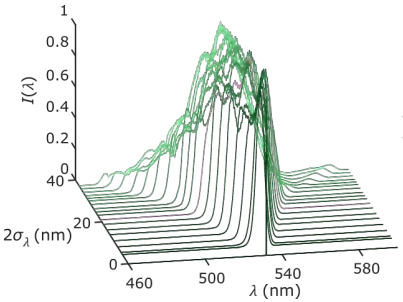
Applications

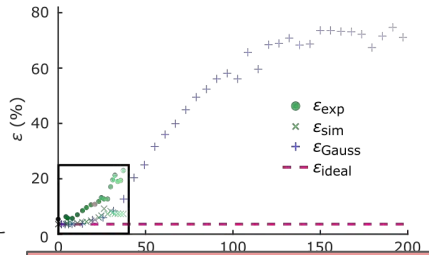
- Fully integrated on-chip inference
  arXiv:2203.05466 (2022)
- Computing Across the Internet's Edge: *Tops/sec @mW on edge devices*
  ArXiv:cs.ET cs.LG 2205.09103 (2022); to appear in Science
- Single-shot NNs: *20,000 MACs / 20 fs: Exascale computing*
  L Bernstein et al, ArXiv:2205.09103 (2022)

Single-shot ONN









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

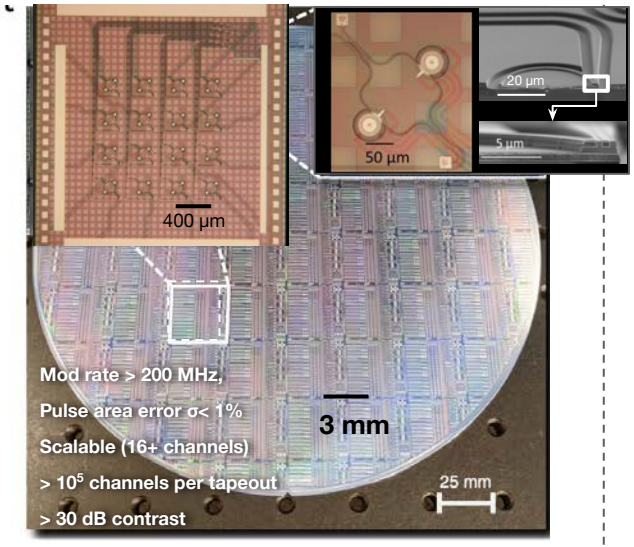
.. but hold on! that's a single query, but how many queries per second?! For this we need fast 2D transmitters..

Liane Bernstein, Alexander Sludds, Christopher Panuski, Sivan Trajtenberg Mills, Ryan Hamerly, and Dirk Englund, Scalable Ultralow Latency Photonic Tensor Processor, CLEO, IEEE, May 19, 2022, STh5G.7.

Answer:

The arrival of scalable control by analog photonic integrated circuits (APICs)

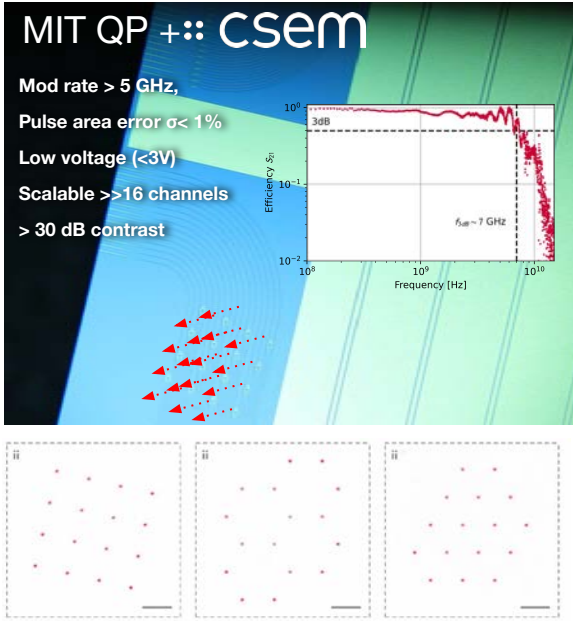
16-CH APIC[2] in 200 mm CMOS fabrication process[1]



[1]Mark Dong et. al . 2021. Nature Phot. (2021)
[2]Adrian Menssen, Artur Hermans, et. al. 2022 CLEO (2022); to be published (2022)

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

16-CH APIC lithium niobate on insulator[3]

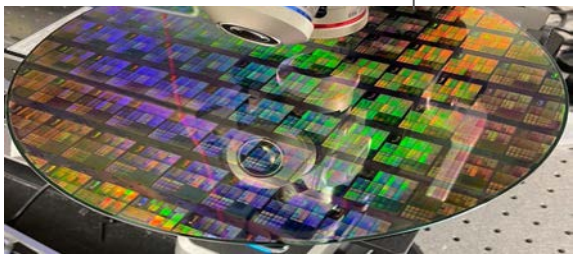


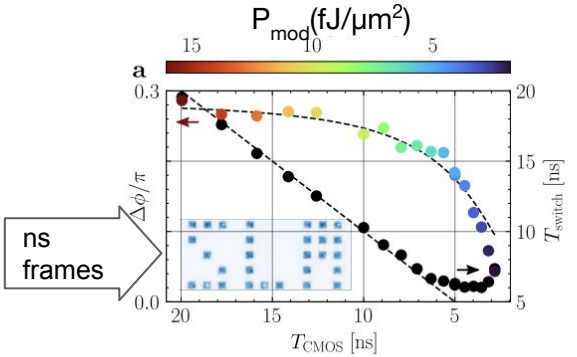
16 programmable projection fields, 16 x 10 GHz bandwidth

[3]Ian Christen et al, CLEO (2022), to be published

2D Spatial Light Modulator with ns switching rate; Si demonstrator [4]:

10 B modulators per wafer





[4] **C. Panuski et al, arXiv:2204.10302 (2022)**

Outline

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

Intro

- A Renaissance in Physics of Computing ?

Hardware

- Programmable photonic integrated circuits (PICs)
  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)
  Hamerly et al., arXiv:2106.03249 (2021)
  R Hamerly, S Bandyopadhyay, DE, arXiv: 2109.05367 (2021)
- EC in ML in-situ training?

Applications

- Fully integrated on-chip inference
  arXiv:2203.05466 (2022)
- Computing Across the Internet's Edge: *Tops/sec @mW on edge devices*
  ArXiv:cs.ET cs.LG 2205.09103 (2022); to appear in Science
- Single-shot NNs: *20,000 MACs / 20 fs: Exascale computing*
  L Bernstein et al, ArXiv:2205.09103 (2022)
- Microlaser-based Coherent Scalable Efficient deep Learning *~ fJ/OP and 25 TeraOP/(mm^2· s)*
  Z. Chen et al, CLEO 2022
  R. Davis et al, CLEO 2022

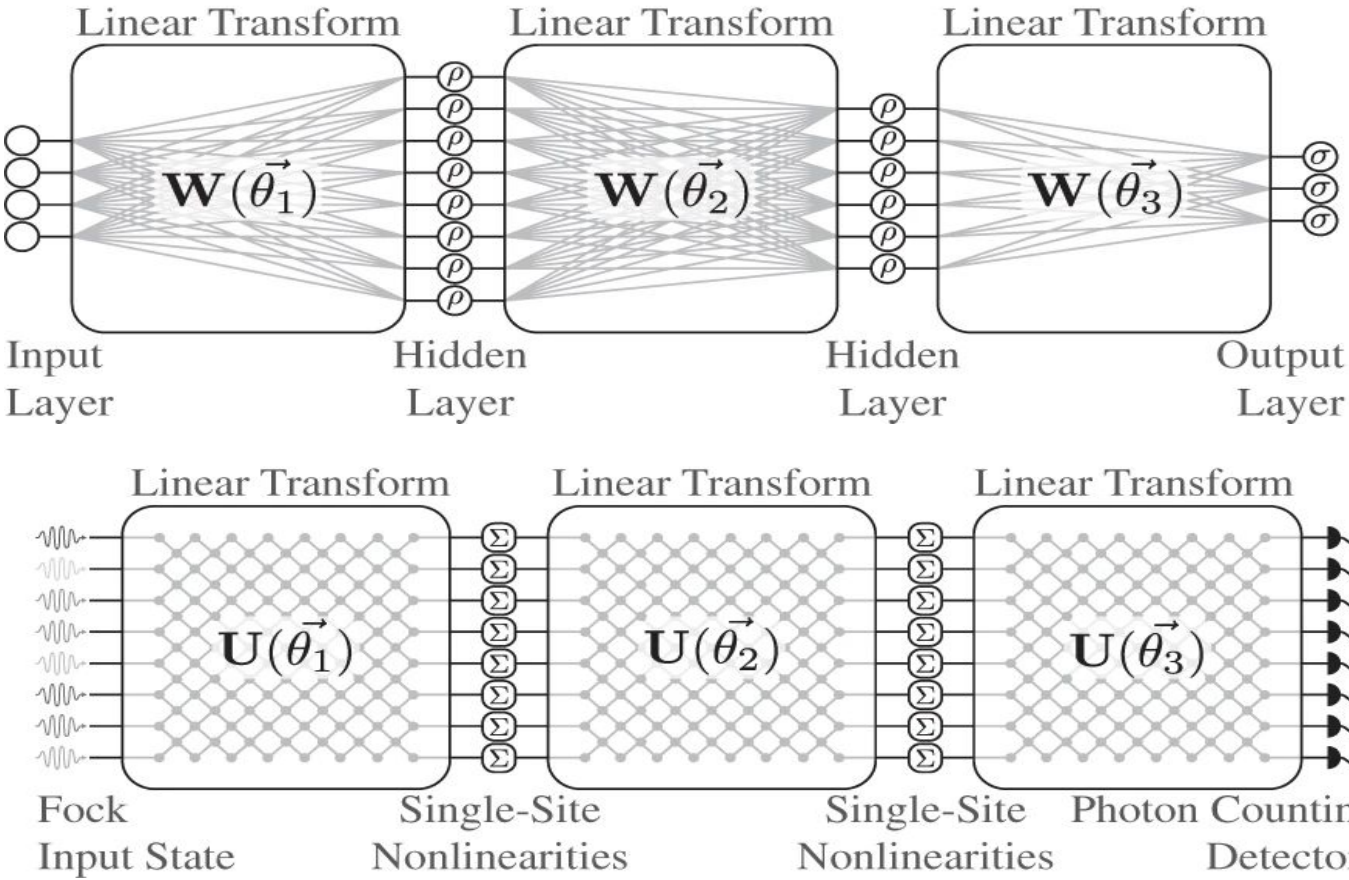
- Quantum optical neural networks: signal processing on quantum states

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

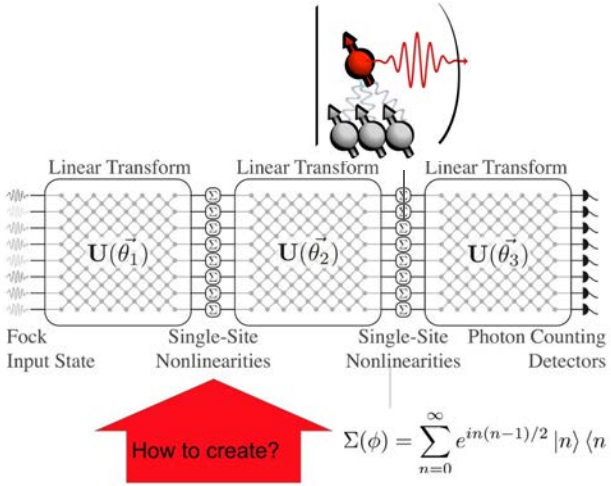
MIT Quantum Photonics

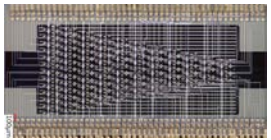
information flow:

physics implementation:



A Quantum Optical Neural Network (QONN)





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

Supervised training[1] →

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

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)

Acknowledgements (ML portion)

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

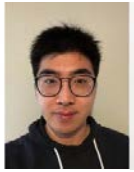
MIT QUANTUM PHOTONICS

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


Mihika Prabhu


Dr. Darius Banundar


Dr. Matthew Streshinsky


Dr. Ari Novack

Collaborators on this work:

MIT: Manya Ghobadi, Joel Emer, Vivienne Sze

MITRE Corp: Gerry Gilbert, Mark Dong, Gen Clark

Sandia NL: M Eichenfield

TU Berlin: T Heuser, N Heermeier, J Lott, Stephan Reitzenstein


Dr. Tom Baehr-Jones


Dr. Euan Allen


Dr. Michael Hochberg


Prof. Manya Ghobadi


Dr. Ryan Hamerly

Funding

AFOSR AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

NSF

NTT

SCIENCE TO SHAPE THE FUTURE OF THE ARMY · ARMY RESEARCH OFFICE · 1951

MITRE

Alexander von Humboldt Stiftung/Foundation