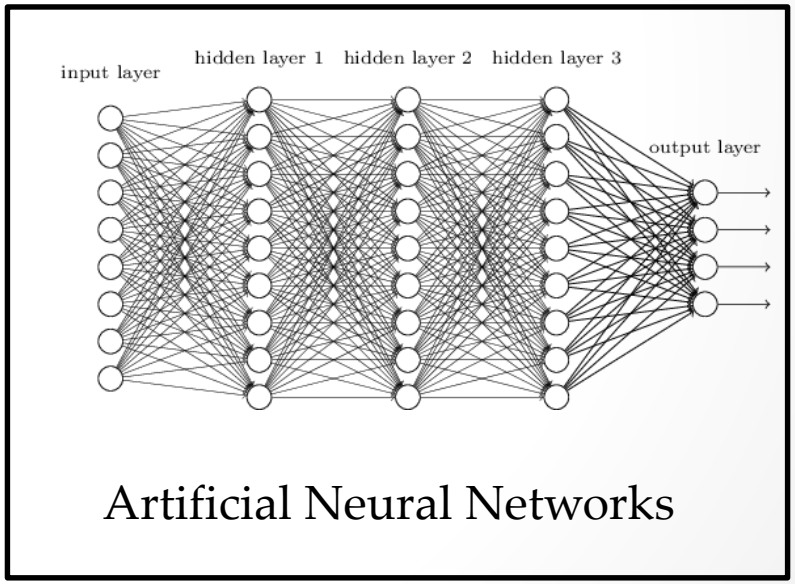
Deep Learning with Coherent Nanophotonic Circuits

Yichen Shen, Nicholas Harris, Dirk Englund, Marin Soljacic
Massachusetts Institute of Technology
@ Berkeley, Oct. 2017

Neuromorphic Computing

Biological Neural Networks



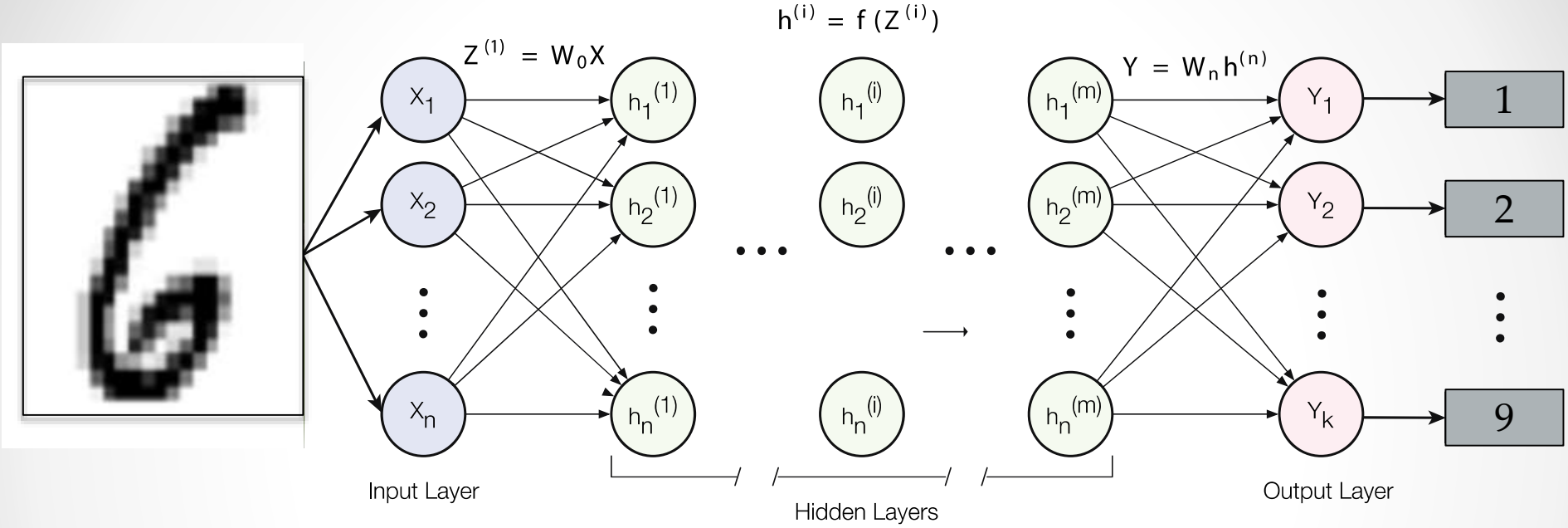
Artificial Neural Networks

Artificial Neural Networks (ANN)

Breakthroughs in deep learning:

- Natural Language Processing (NLP)
- Game Playing (Go, Atari)
- Autonomous Vehicles
- Control
- Ad Placement
- Researches (drug discovery, material study)
- Etc.

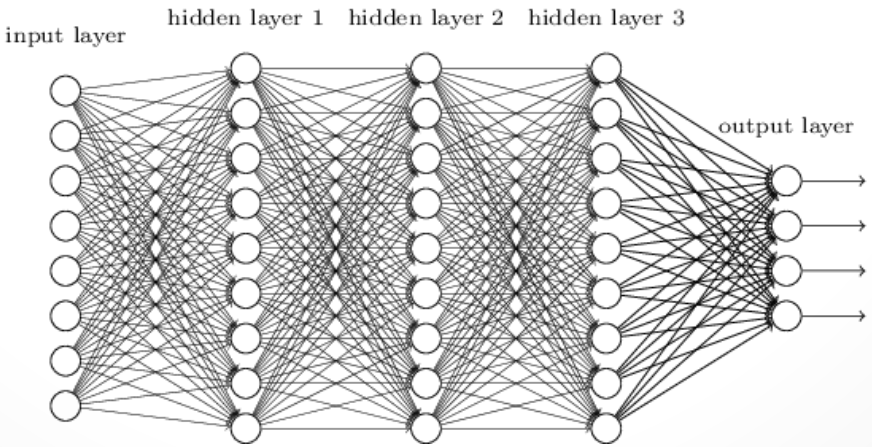
Basic Algorithm of ANN



Matrix Multiplication: $z_j^{(1)} = \sum_{i=1}^{n} w_{ji}^{(1)} x_i \quad \cdots \quad z_j^{(k)} = \sum_{i=1}^{n} w_{ji}^{(k)} h_i^{(k)} \quad \cdots$

Nonlinear Activation: $h_j^{(1)} = f(z_j^{(1)}) \quad \cdots \quad h_j^{(k)} = f(z_j^{(k)}) \quad \cdots$

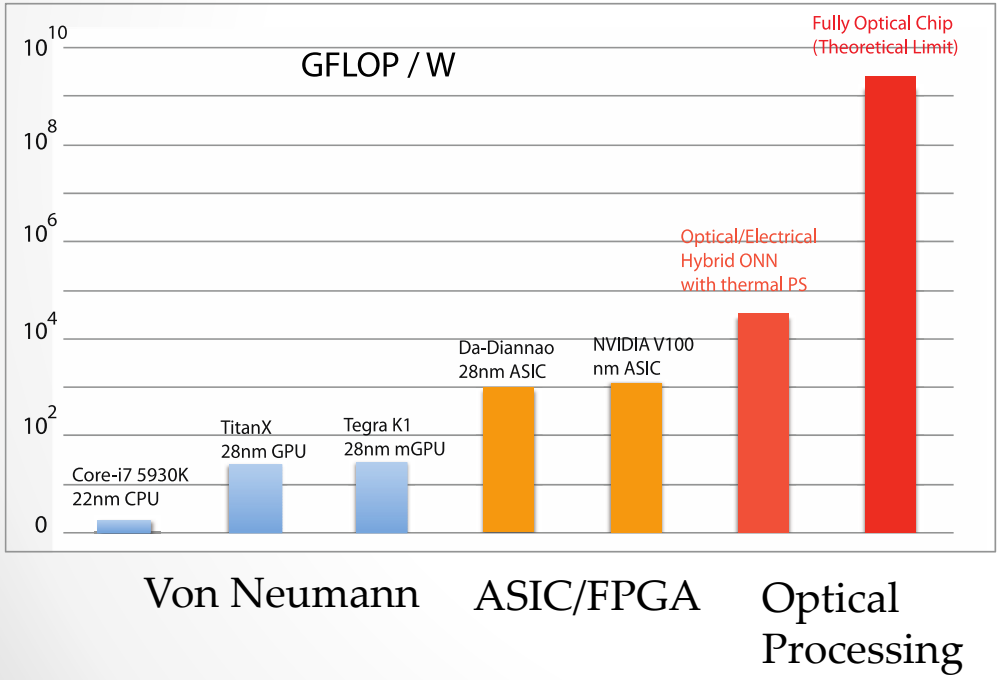
Hardware and Data Enable Deep Learning



The Need for Speed

More Data → Bigger Models → More need for Computation

But Moore's Law is no-longer providing more computation…



The Market:

On clouds:
Millions of high power AI processors ($10,000 each) in data centers by 2020

On premise:
Billions of compact AI processors needed due to the rise of autonomouse driving, AR and IoT.

Optical AI Computing

The New York Times

Science

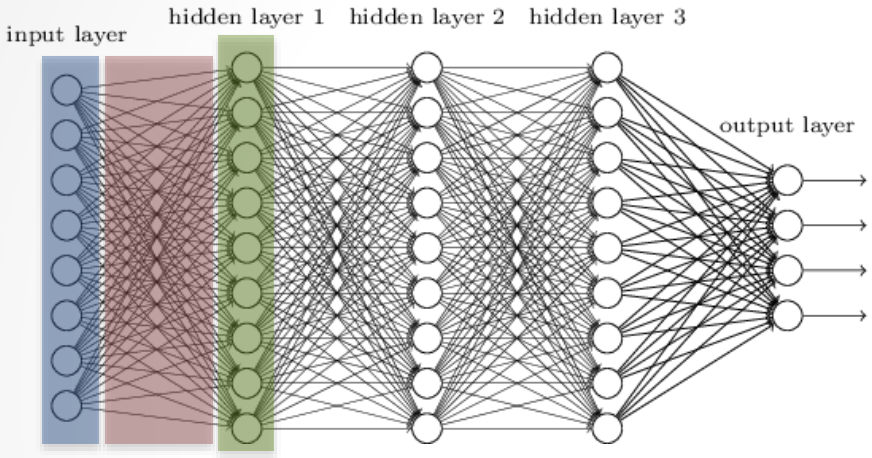
LIGHT MAY BE KEY TO NEW GENERATION OF FAST COMPUTERS

By WILLIAM J. BROAD
Published: October 22, 1985

SINCE its start nearly half a century ago, the computer revolution has advanced by virtue of a simple physical phenomenon: that streams of speeding electrons can start or stop the flow of other streams of electrons. In short, electrons can act as a switch.

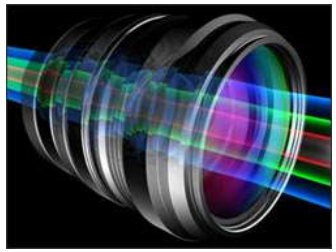
In Deep Learning

Key Operation is dense M x V

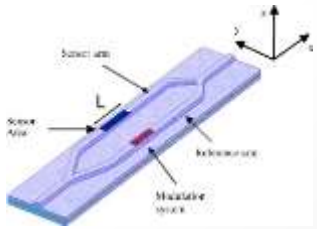


$$b_i = W_{ij} \times a_j$$

In Optics, Matrix Multiplication is very common & (usually) consumes no energy !



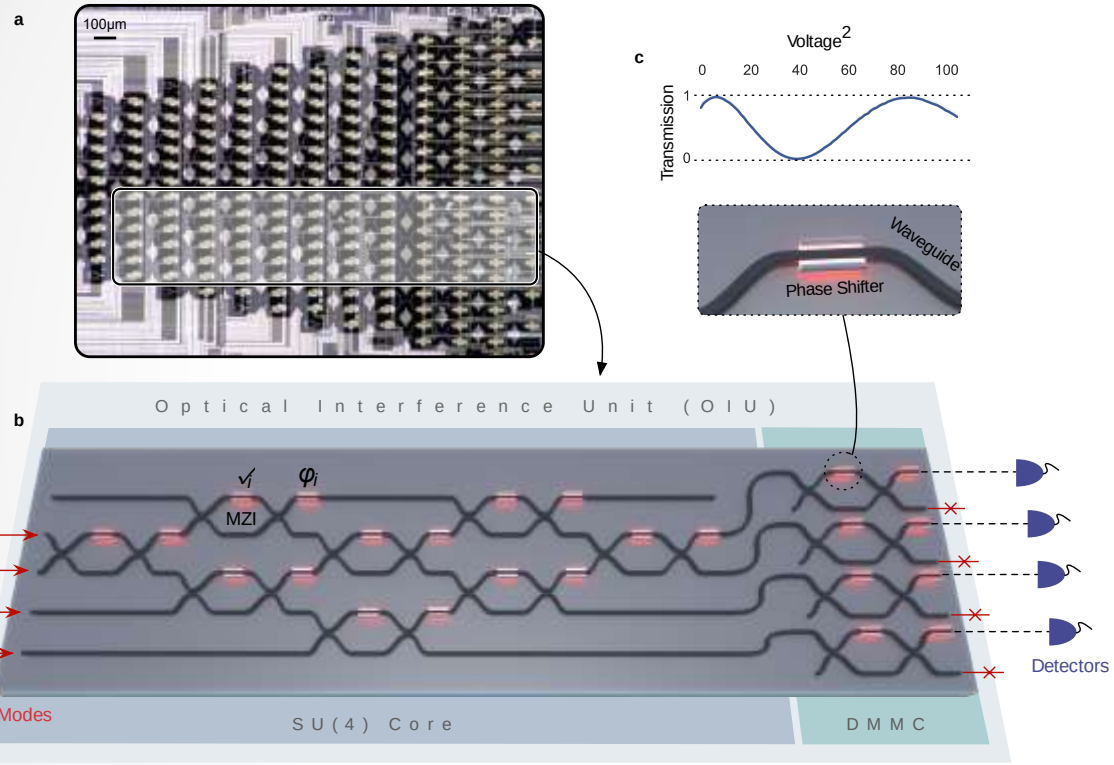
Convolution / FFT

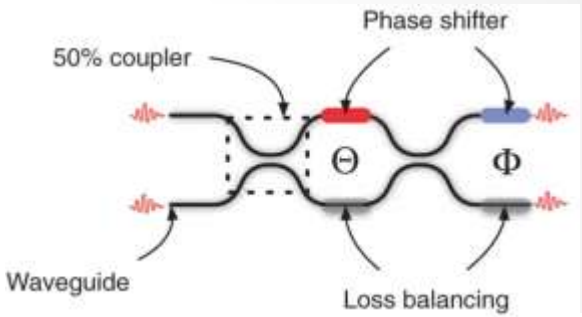


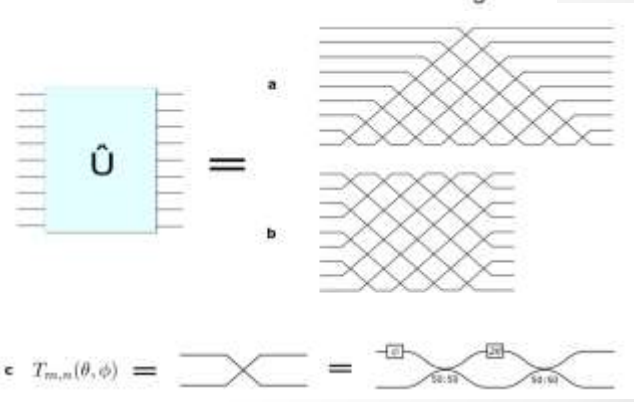
Matrix Multiplication

Programmable Nanophotonic Processors

$$\begin{pmatrix} e^{i\phi}\cos\theta & -e^{i\phi}\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$







J. Mower *et al*, Physical Reviews A, 92, 032322 (2015)
Carolan, Jacques, et al. "Universal linear optics." Science 349.6249 (2015): 711-716.

ANN does NOT require high resolution

Category	Method	Weights (# of bits)	Activations (# of bits)	Accuracy Loss vs. 32-bit float (%)
Dynamic Fixed Point	w/o fine-tuning	8	10	0.4
	w/ fine-tuning	8	8	0.6
Reduce weight	Ternary weights Networks (TWN)	2*	32	3.7
	Trained Ternary Quantization (TTQ)	2*	32	0.6
	Binary Connect (BC)	1	32	19.2
	Binary Weight Net (BWN)	1*	32	0.8
Reduce weight and activation	Binarized Neural Net (BNN)	1	1	29.8
	XNOR-Net	1*	1	11
Non-Linear	LogNet	5(conv), 4(fc)	4	3.2
	Weight Sharing	8(conv), 4(fc)	16	0

* first and last layers are 32-bit float

Sze et al, arXiv:1703.09039 (2017)

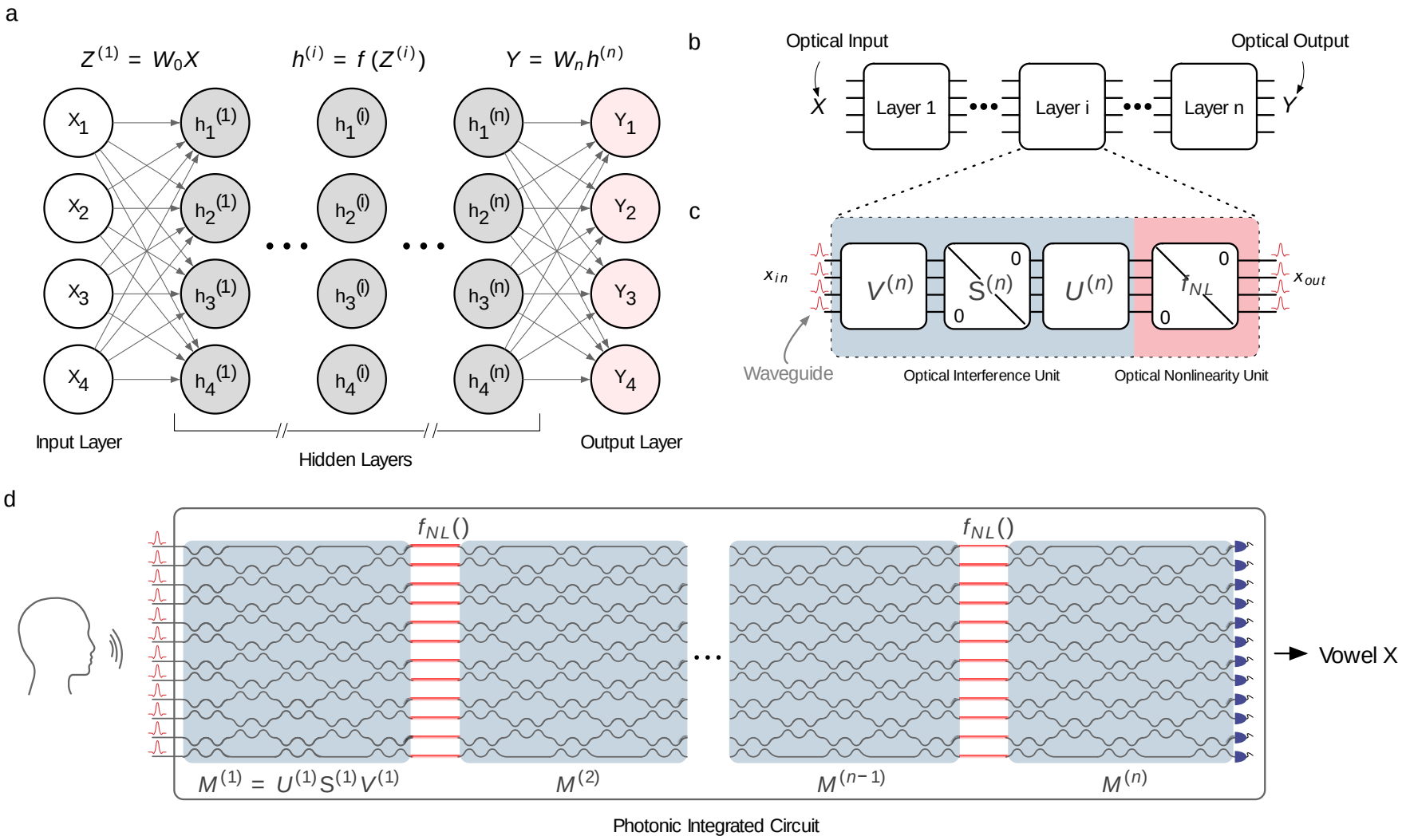
Deep Learning Inference is "Passive"

Once the Optical Neural Network is trained, no need to update the weights frequently…

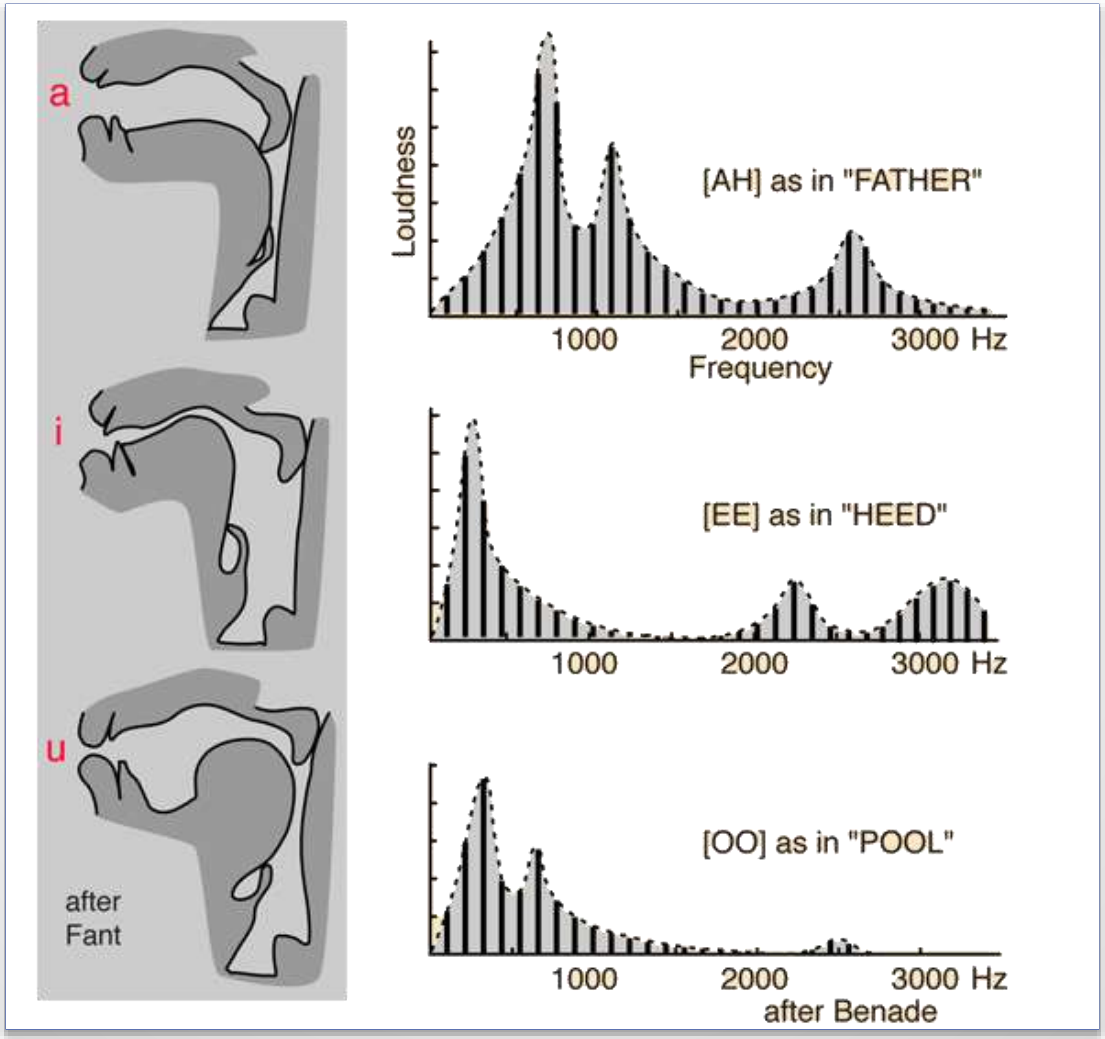
Deep Learning is very parallelizable

Multiple wavelengths can be used to simultaneously execute batch of data

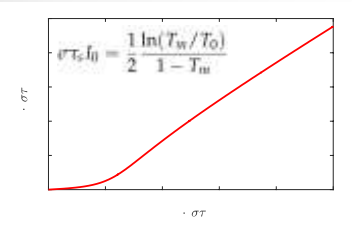
Coherent Optical Neural Networks (ONN)

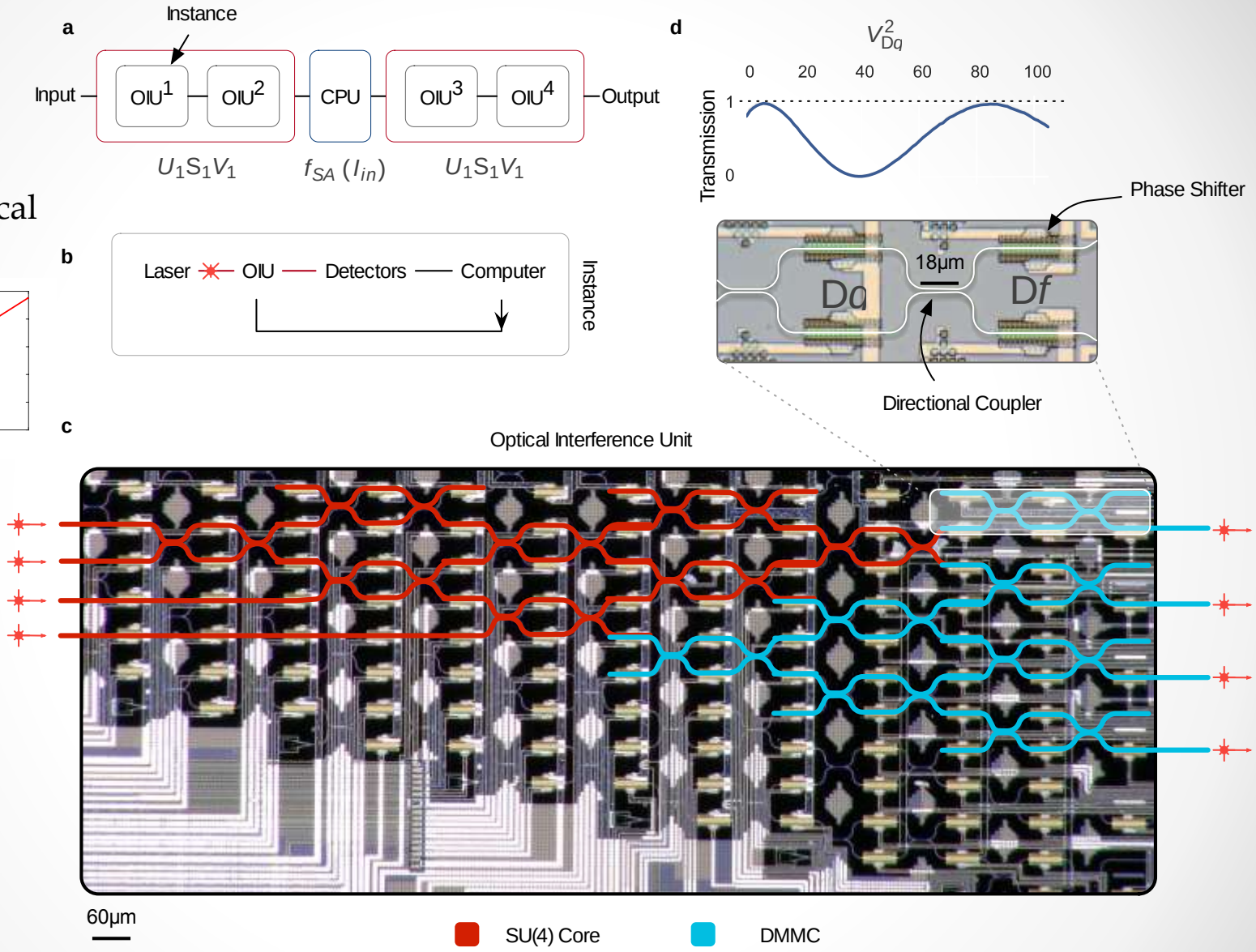


Optical Vowel Recognition (4d 4 classes)



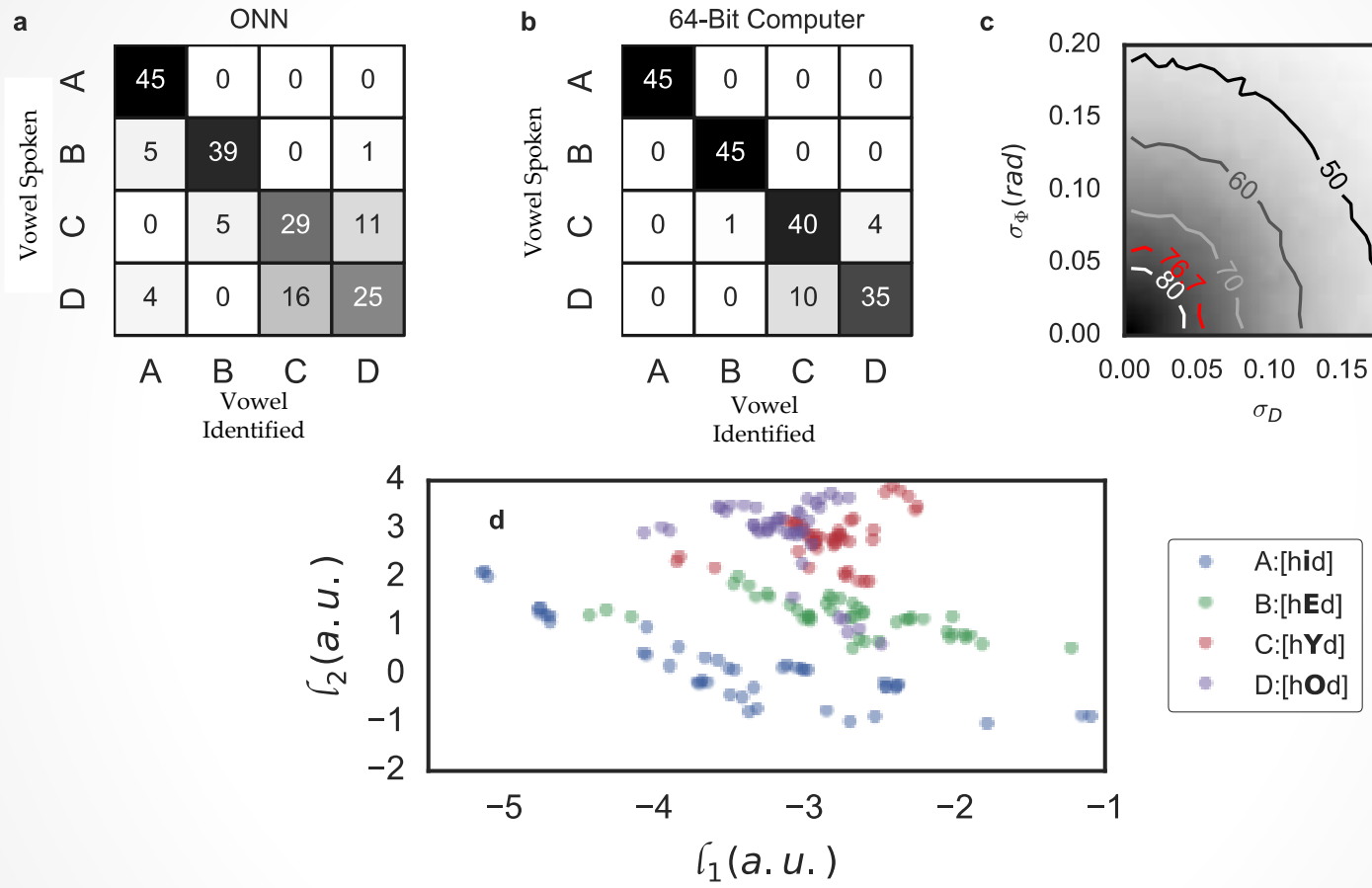
Simulated Optical Nonlinearity





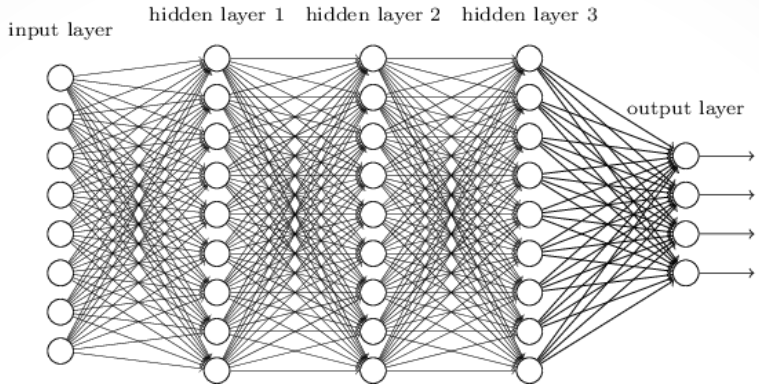
Y. Shen and N. Harris et al, Nature Photonics, 11, 441-446 (2017)

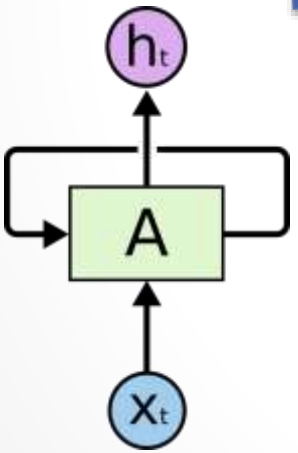
Experimental Result

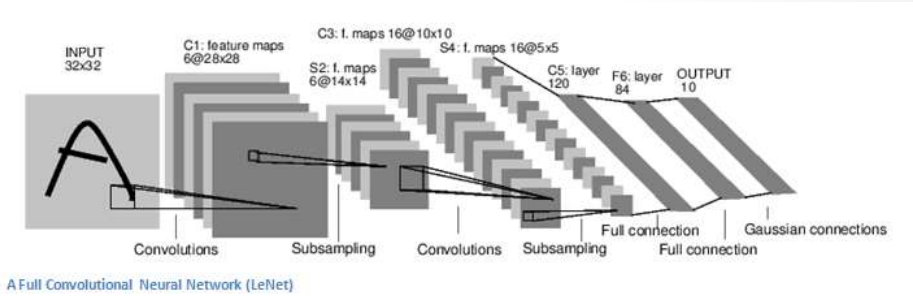


Simulation Result: 165/180=91.7%
Experiment Result: 138/180=76.7%

Fully Connected Neural Networks







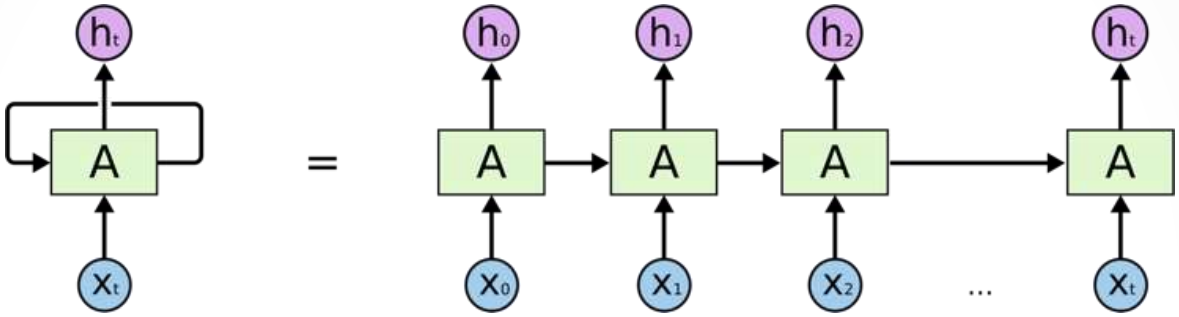
A Full Convolutional Neural Network (LeNet)

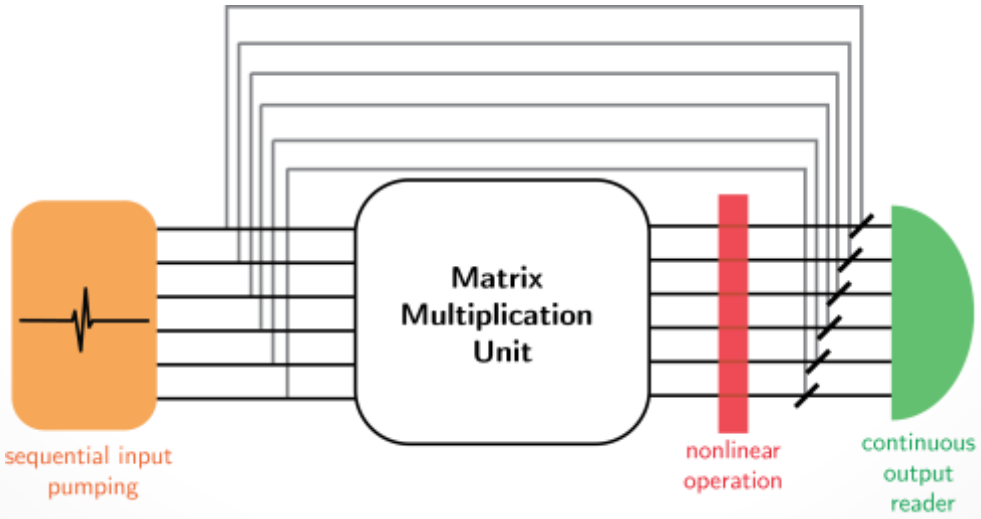
Recurrent Neural Networks

Convolutional Neural Networks

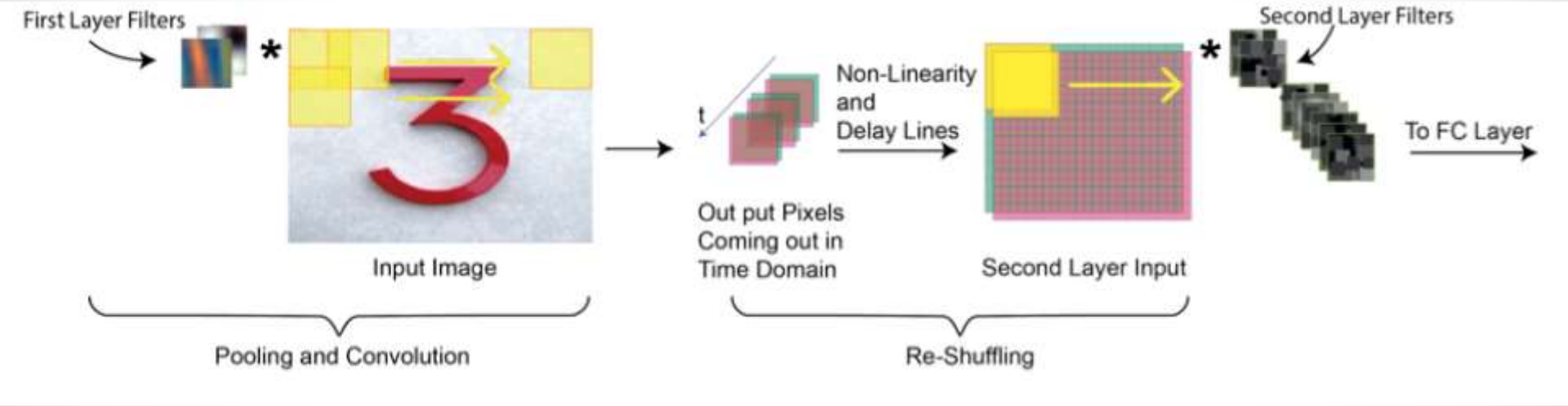
Recurrent Neural Networks

Commonly used for Speech Recognition and Language Processing



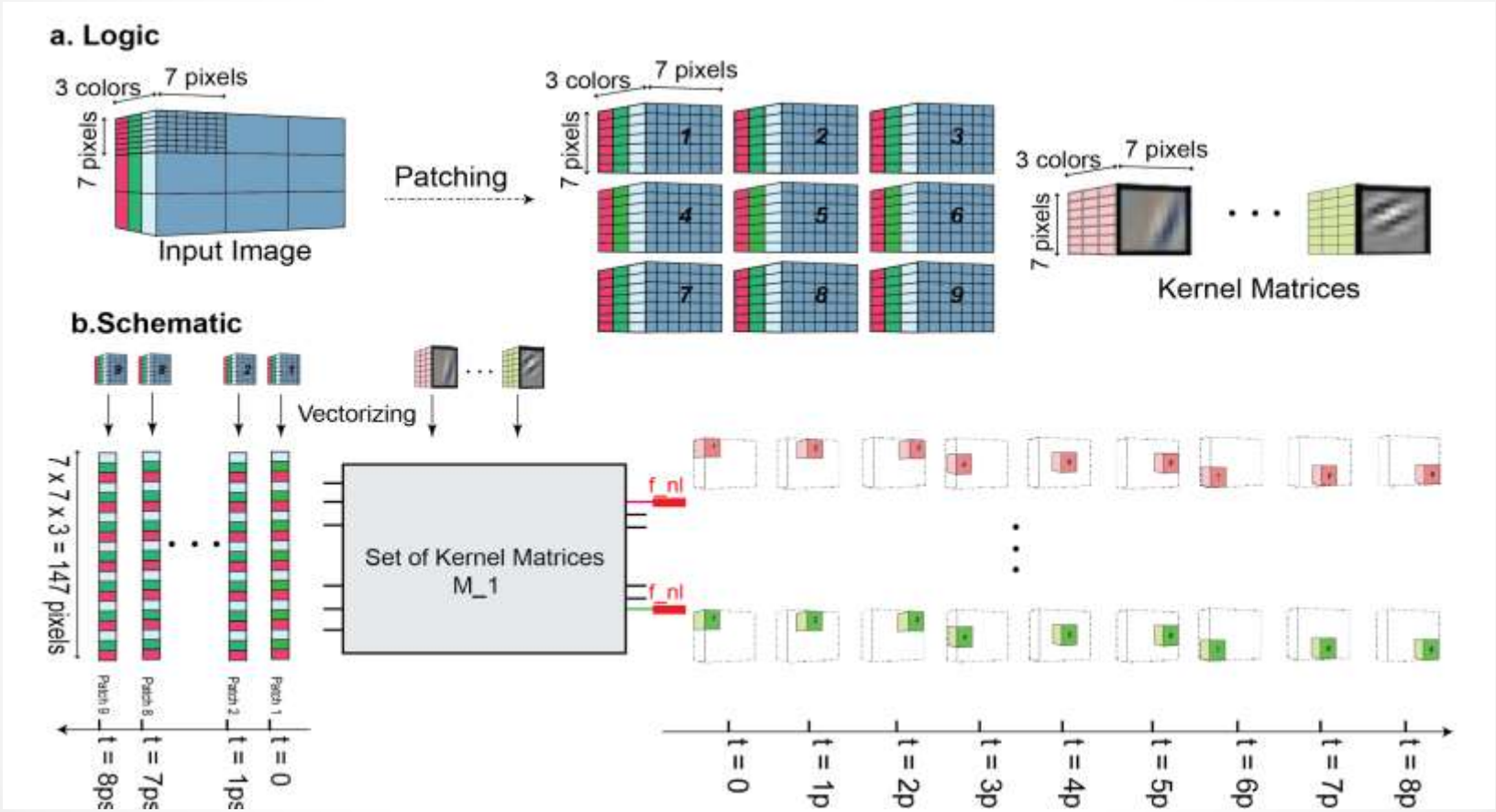


Convolution Neural Networks

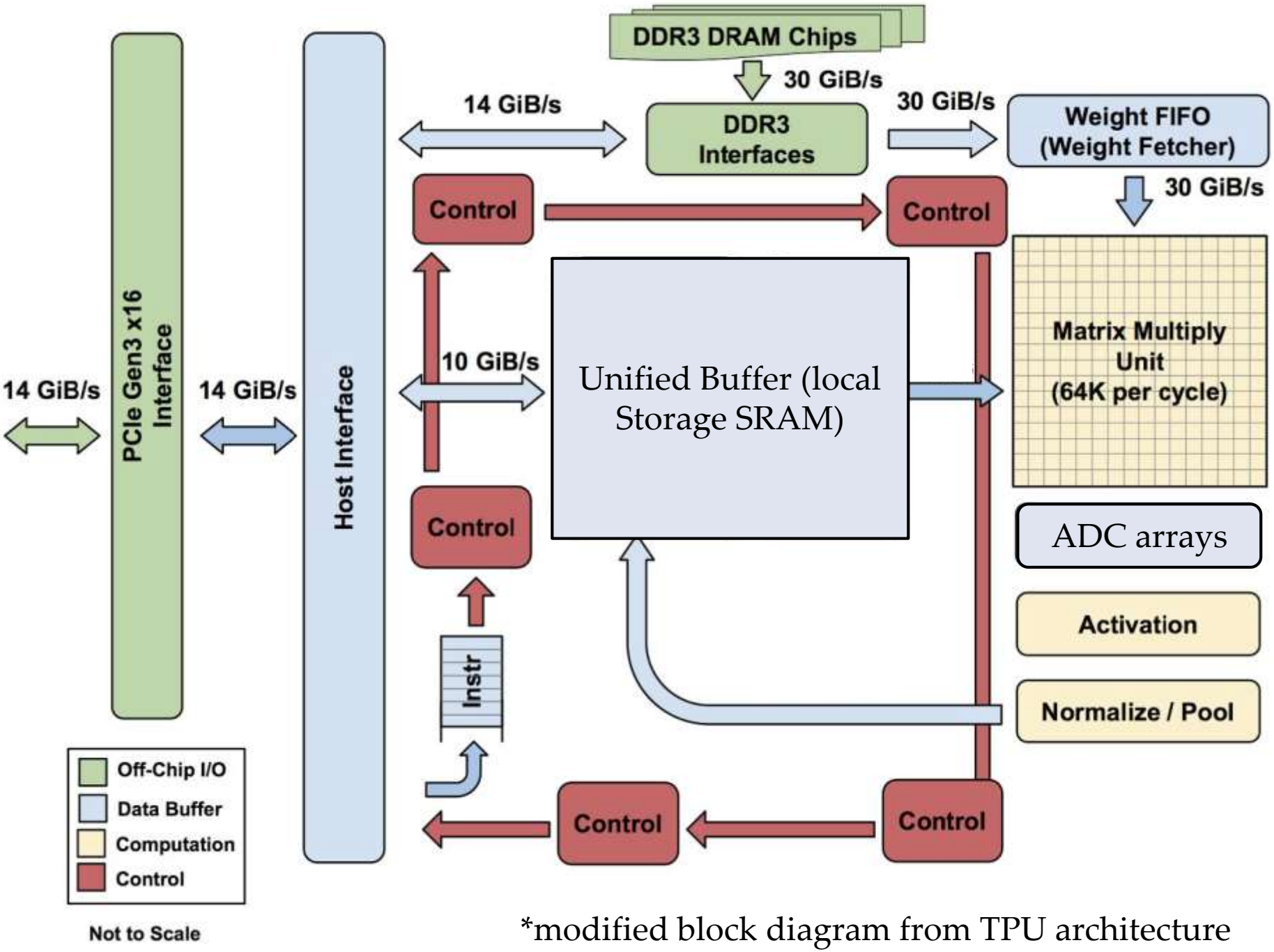


Scott Skirlo and Yichen Shen et al, Manuscript in Preparation

Optical Convolutional Neural Network



Scott Skirlo and Yichen Shen et al, Manuscript in Preparation



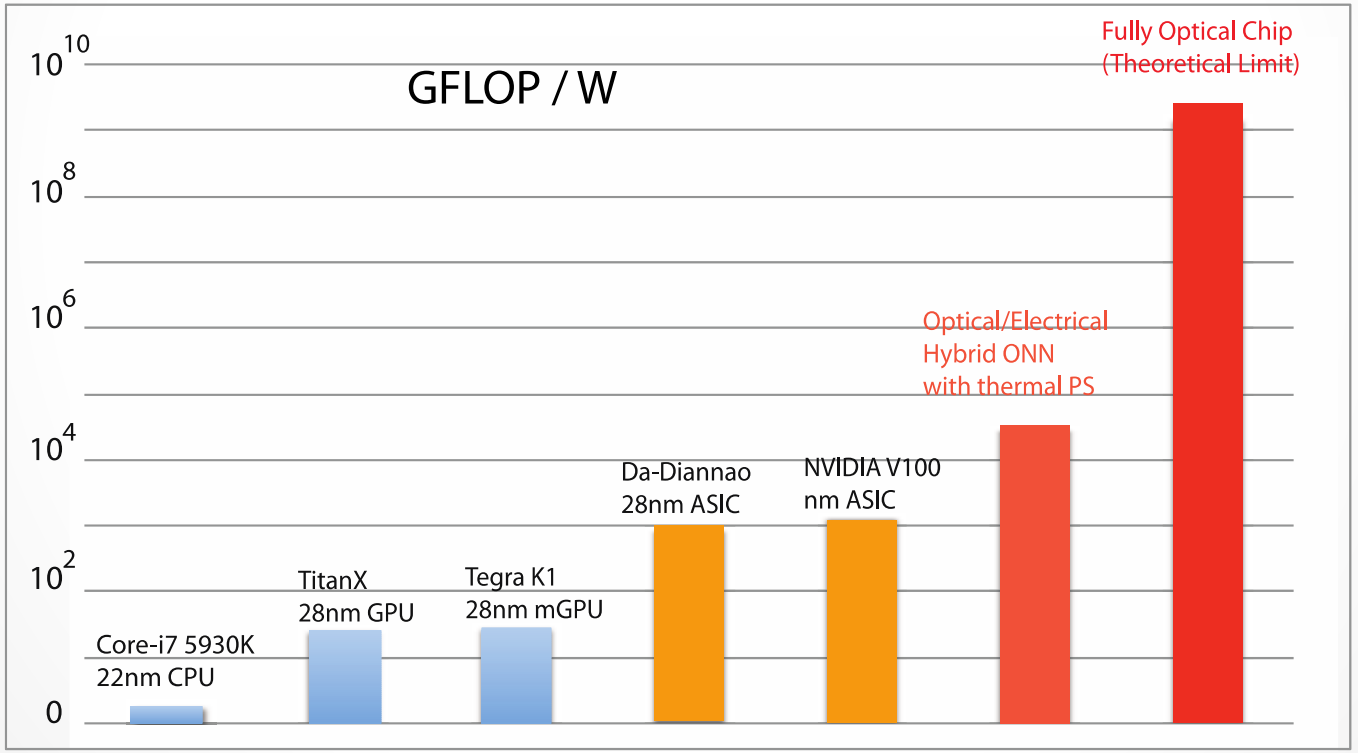
*modified block diagram from TPU architecture

Speed and Energy Efficiency Comparison with Electrical ANN

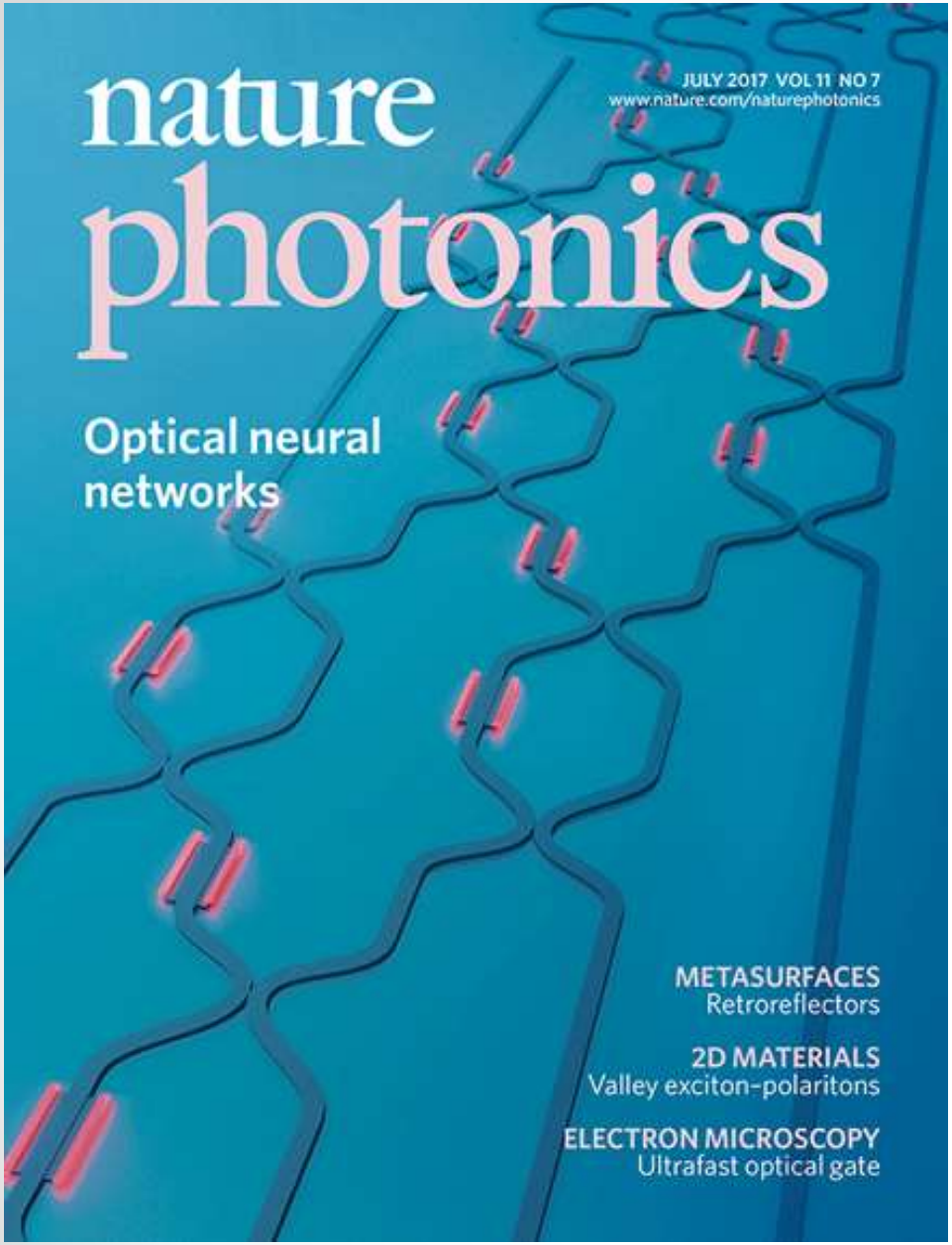
	NVIDIA TITAN X	ONN (with thermal PS)
Architecture	Von Neumann	Neuromorphic
Power Consumption	1 kW	1-2 kW
Operation Speed	10 TFLOP	10,000 TFLOP

$$P = P_{PS} + P_{OE/EO}$$
$$= 1mW \times N^2 + 300mW \times 2N$$
$$= 1.6\ \text{kW}$$

$$Speed = 10bit \times N^2 \times 1\ GHz$$
$$= 10{,}000\ TFLOP/s$$



Y. Shen and N. Harris et al, Nature Photonics 11, 441 (2017)


nature photonics
JULY 2017 VOL 11 NO 7
www.nature.com/naturephotonics
Optical neural networks
METASURFACES
Retroreflectors
2D MATERIALS
Valley exciton-polaritons
ELECTRON MICROSCOPY
Ultrafast optical gate


Today's Spotlight
Learning with light


SCIENTIFIC AMERICAN
COMPUTING
Light-Powered Computers Brighten AI's Future
Optical computers may have finally found a use—improving artificial intelligence

Some History on Optical Neural Networks

2005

"*The biggest issue with this paper is that it relies on neural networks.*"

Anonymous Reviewer

2016

Springtime for AI: The Rise of Deep Learning

After decades of disappointment, artificial intelligence is finally catching up to its early promise, thanks to a powerful technique called deep learning

By Yoshua Bengio on June 1, 2016

SCIENTIFIC AMERICAN.

Acknowledgement

Nicholas Harris
PhD, EECS
MIT

Scott Skirlo
PhD, Physics
MIT

Li Jing
PhD, Physics
MIT

Massachusetts Institute of Technology

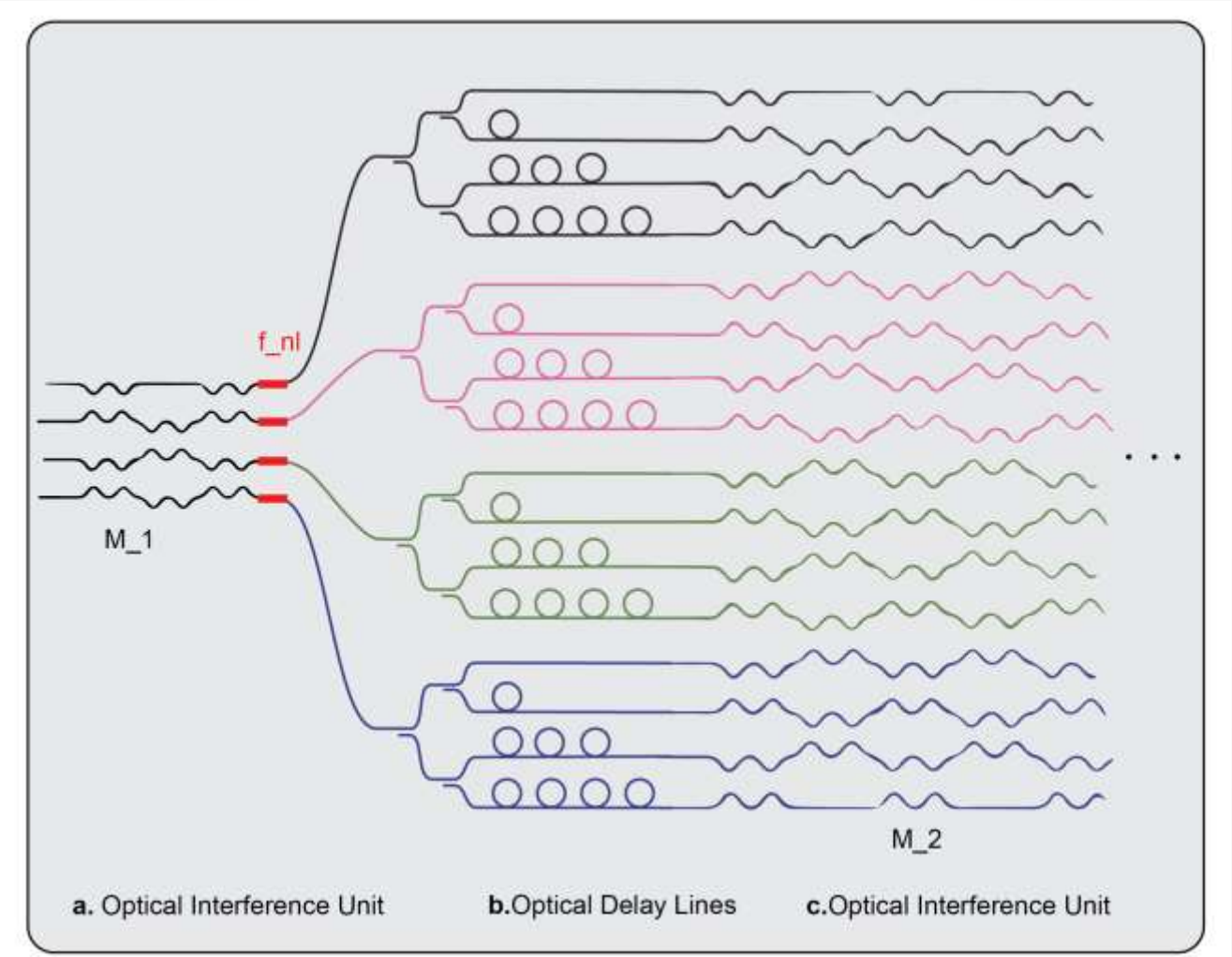
Prof. Dirk Englund
EECS, MIT

Prof. Marin Soljacic
Physics, MIT

Institute for Soldier Nanotechnologies

U.S. DEPARTMENT OF ENERGY

Optical Convolutional Neural Network

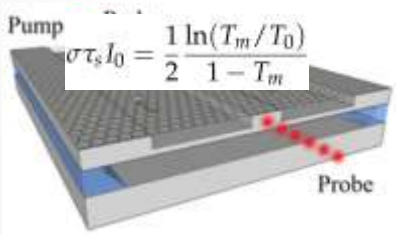


Scott Skirlo and Yichen Shen et al, Manuscript in Preparation

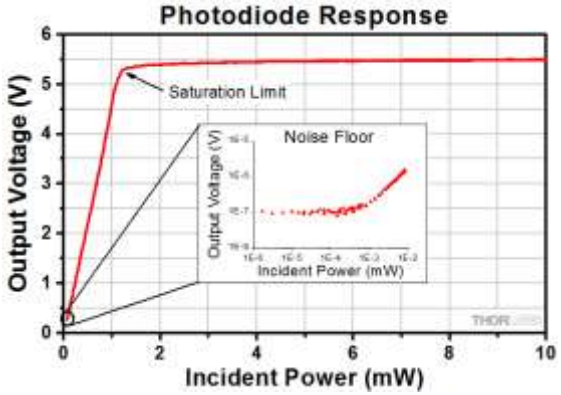
Nonlinearity

For Deep Learning, the constraint on nonlinearity is **weak**

Saturable Absorption



Photodiode



A. Selden, *British Journal of Applied Physics* 18 , 743 (1967)
M. Soljacic, *Physical Review E* 66, 055601 (2002)
Z. Cheng et al, I*EEE Journal of Selected Topics in Quantum Electronics* 20.1 (2014): 43-48.