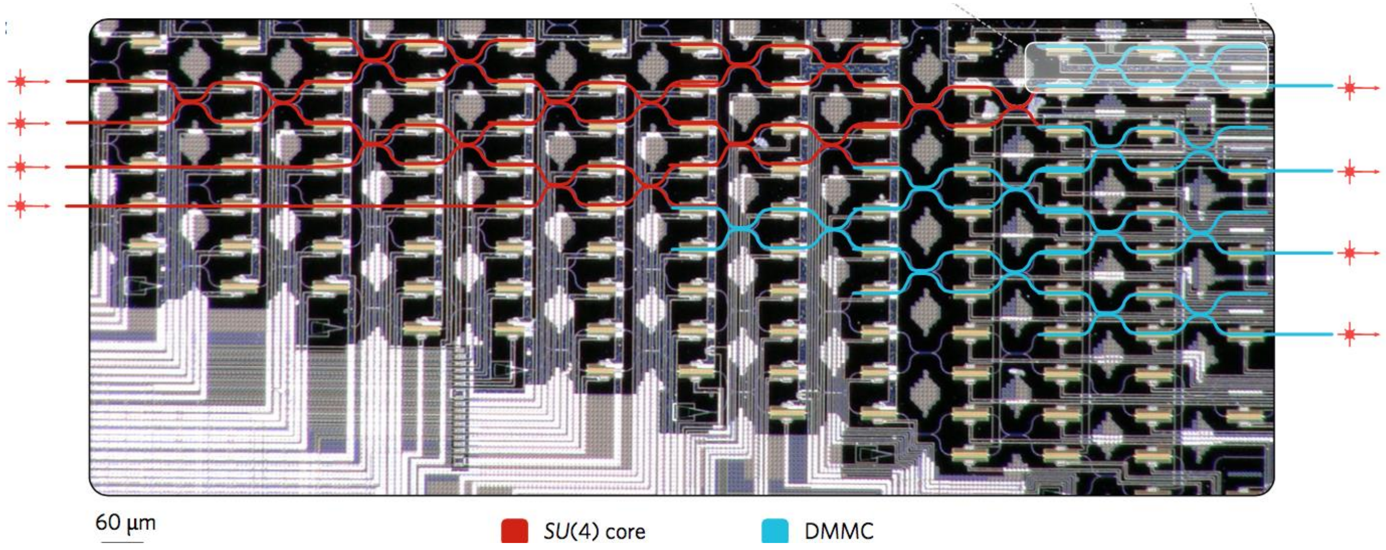
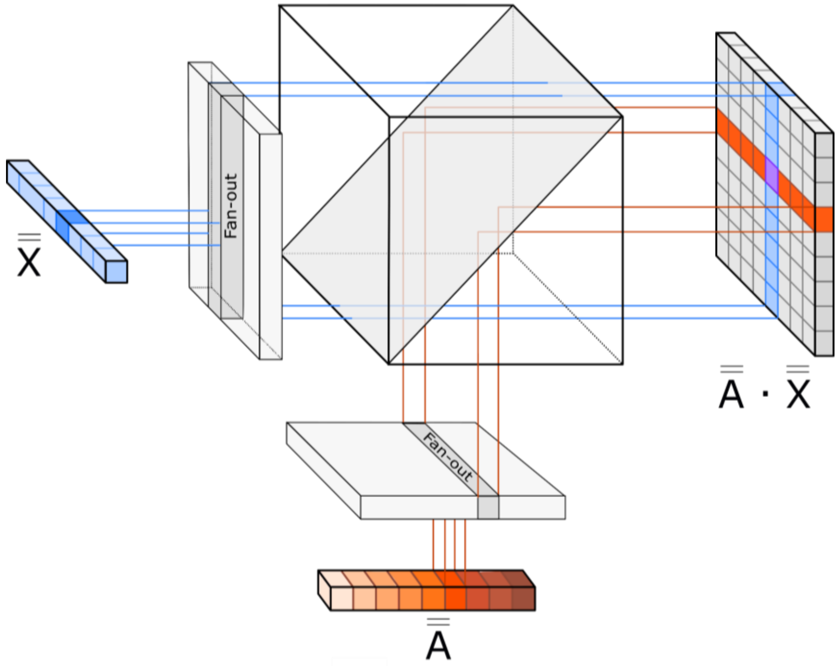
Neuromorphic Photonics

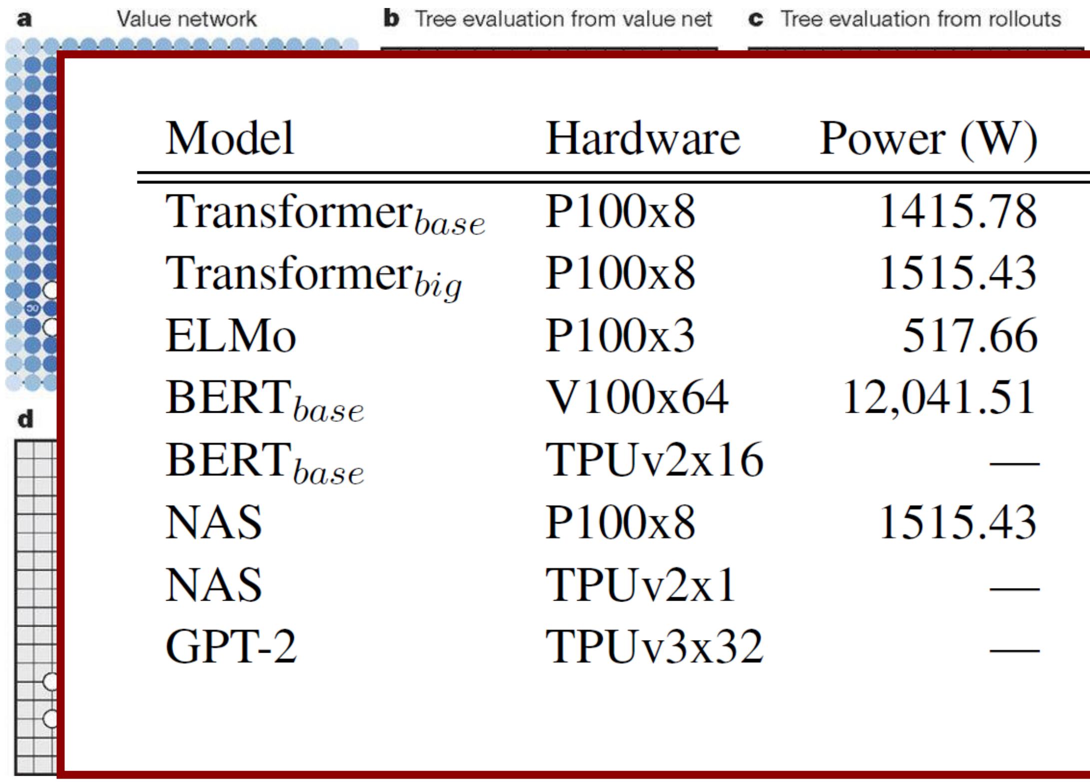
Liane Bernstein, Alexander Sludds, Ryan Hamerly, Dirk Englund
Quantum Photonics Group, MIT Department of EECS

Fast Machine Learning
Fermilab LPC, Batavia, IL
September 11, 2019



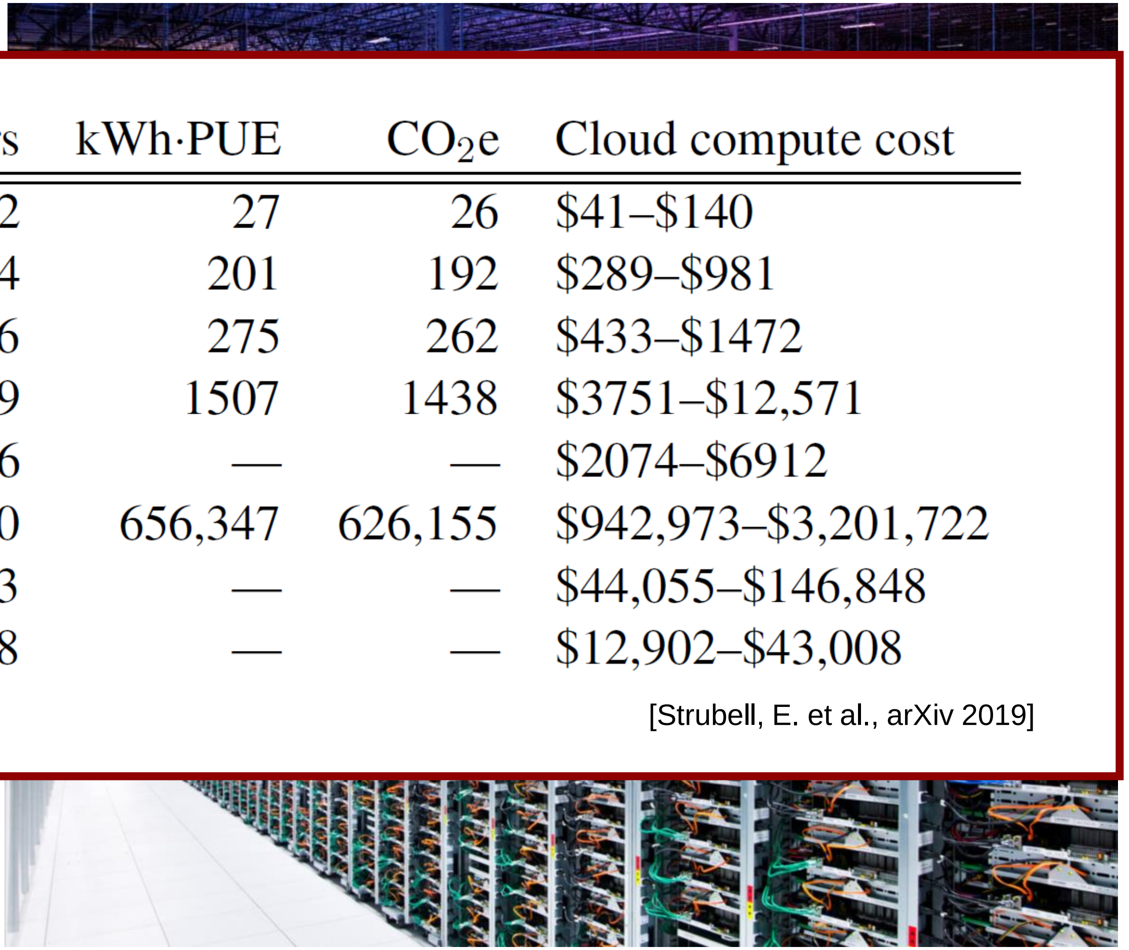


Deep Neural Networks: Current Limitations



Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

[Strubell, E. et al., arXiv 2019]



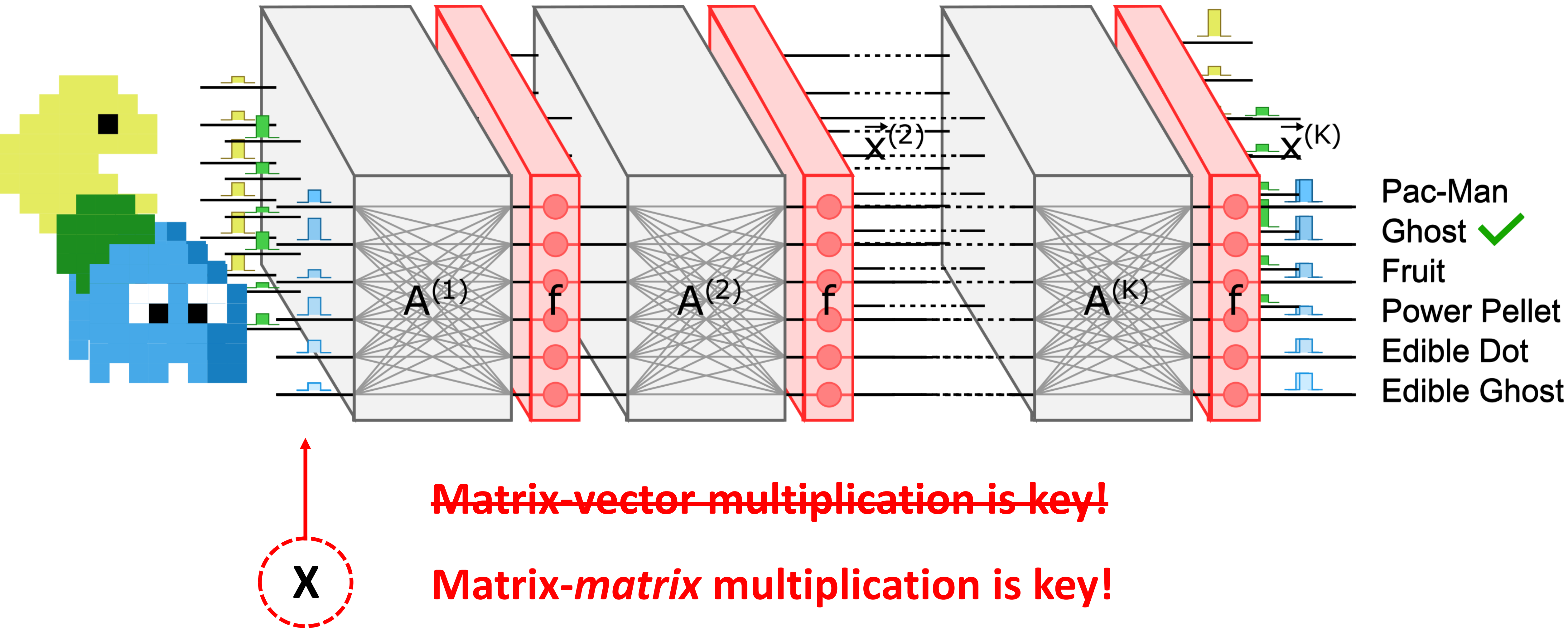
- Training: 3 weeks for 340 million training steps
- During gameplay: 1,920 CPUs, and 280 GPUs

[google.com/about/datacenters/gallery/]

[Sliver, D., *Nature* 2016]

Energy consumption and latency

Deep Neural Networks: Inference



~~Matrix-vector multiplication is key!~~

Matrix-*matrix* multiplication is key!

Energy Consumption in DNNs

Operations required for matrix multiplication:

- Multiplication
- Addition

$\}$ Multiply-accumulate (MAC)

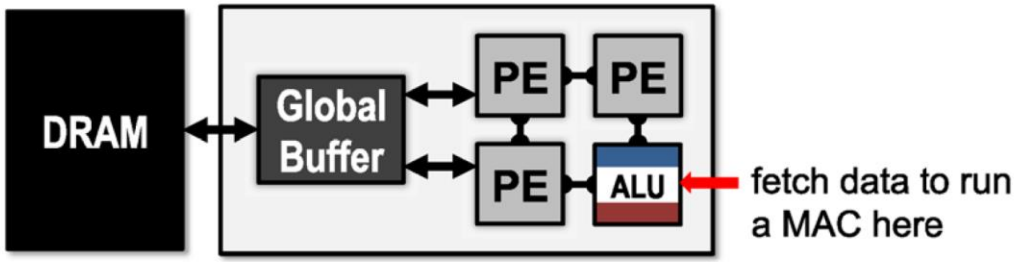
- **Memory access**

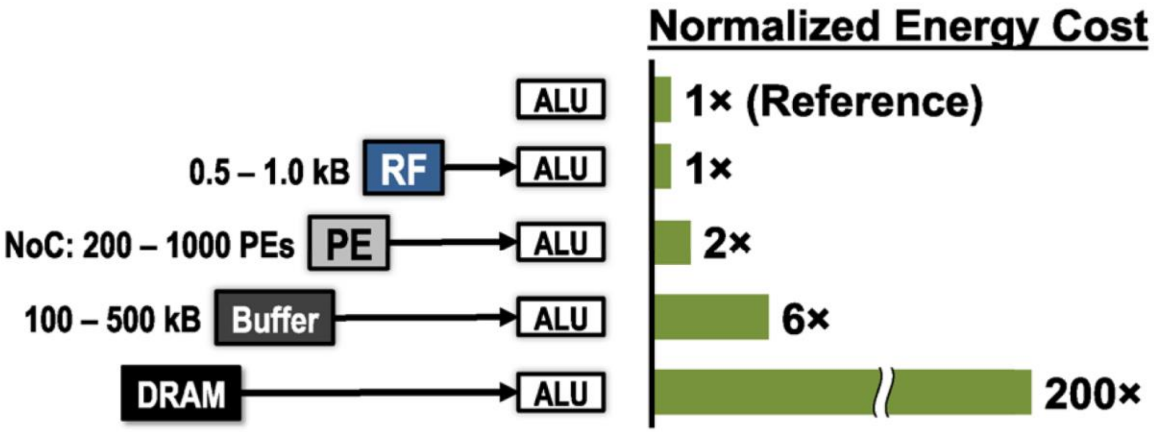
$$Y = A \cdot X = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix}$$

DRAM access consumes ~100-1000x more energy than the MAC

Specialized Hardware for DNNs

- GPU, TPU, ASICs (Eyeriss*), etc.
- Optimized memory hierarchy for data reuse in matrix mult. & conv.





*[Sze, V. et al., *Proc. IEEE* 2017]

Hardware optimized for DNNs/CNNs with 256 processing elements

State-of-the-art chips for NNs*:

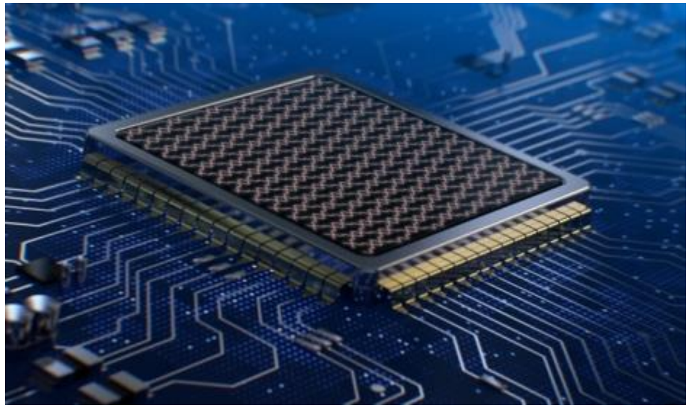
- NVIDIA ~2pJ/MAC
- TPU (Google) ~1pJ/MAC
- Eyeriss ~1pJ/MAC

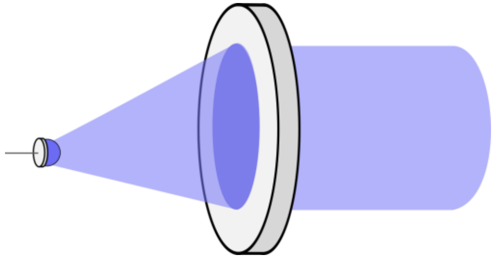
*including memory access

New Paradigm: Photonics

Low energy consumption

- Passive data transmission
- Passive fan-out
- Multiplication

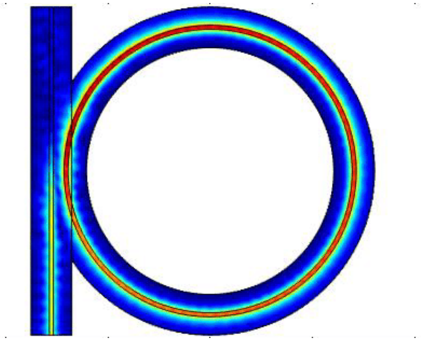




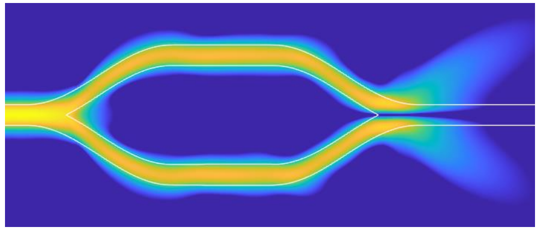
news.mit.edu

High speed

- Modulators operate at 10s of GHz
- Data transmission
  at the speed of light

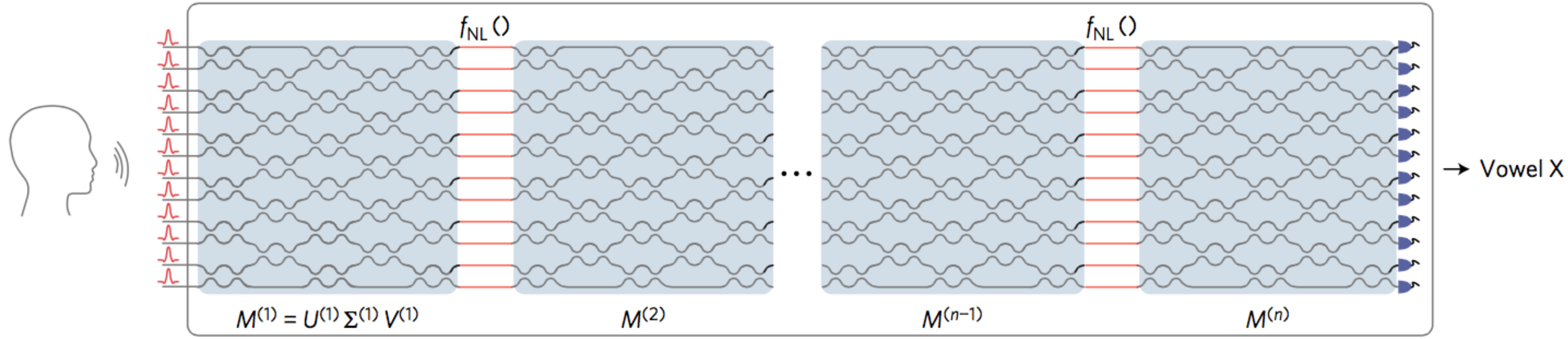


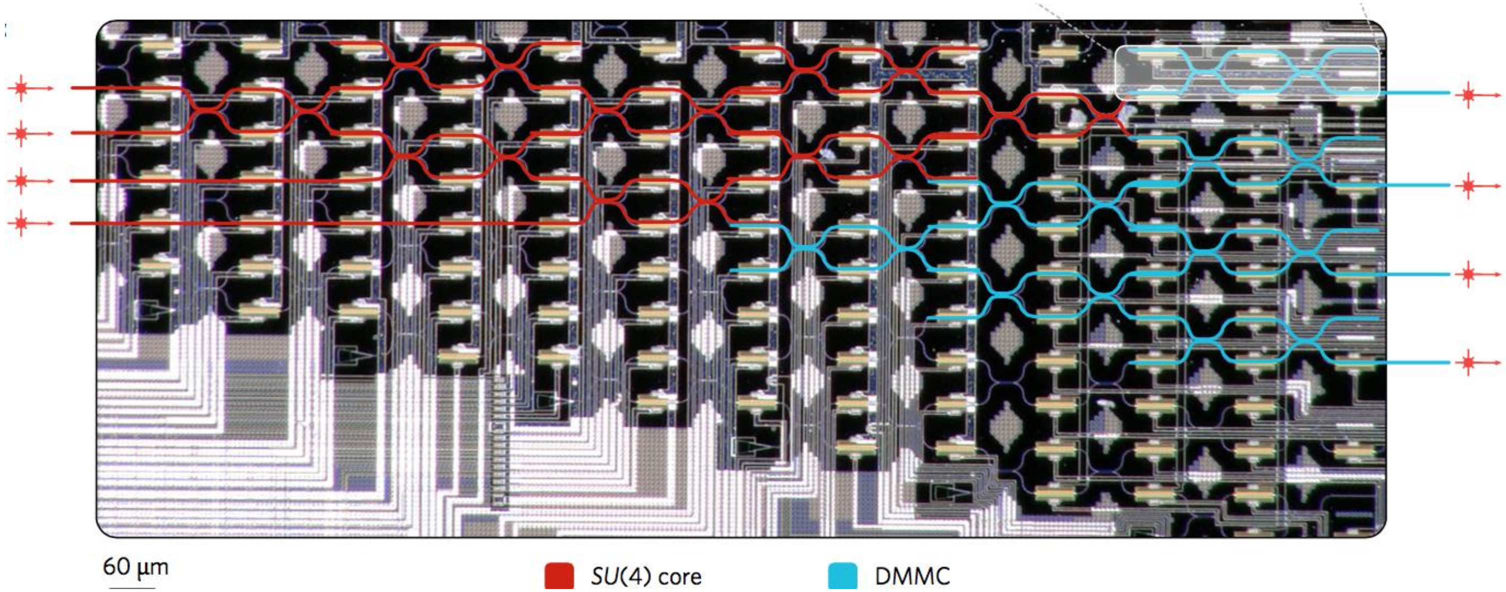
[Xiao, Y. et al., ACP 2018]



codeseeder.com

Past Work: Optical Neural Network (All On-Chip)





- Fast
- Very low energy
- But requires (N^2) phase shifters

↓

Difficult to scale beyond 100s of neurons

[*Shen, Y., *Harris, N.C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., and Soljacic, M., *Nature Photonics* 2017]

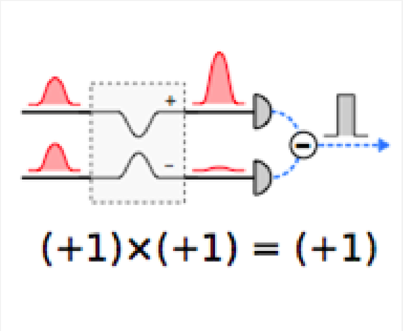
New Approach: HD-ONN

Want to achieve *scalable* ONN with combined advantages of:

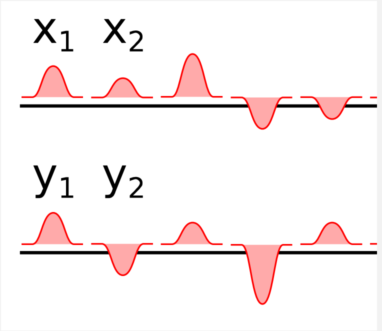
- Free-space optics
- Nanophotonics
- Electronics

Homodyne detection

Photoelectric multiplication



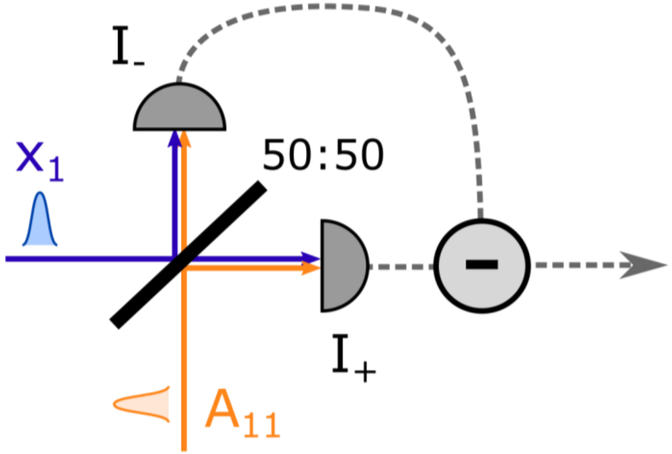
Time multiplexing

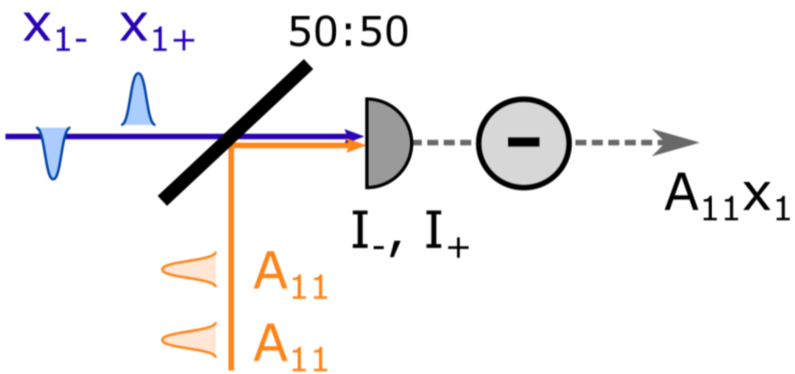


Free-space optical propagation

[Hamerly, R., Bernstein, L., Sludds, A., Soljacic, M., Englund, D., *PRX* 2019]

Homodyne Detection: Optical Multiplication





$$I_+ = \left| -\frac{i}{\sqrt{2}} A_{11} + \frac{1}{\sqrt{2}} x_1 \right|^2$$

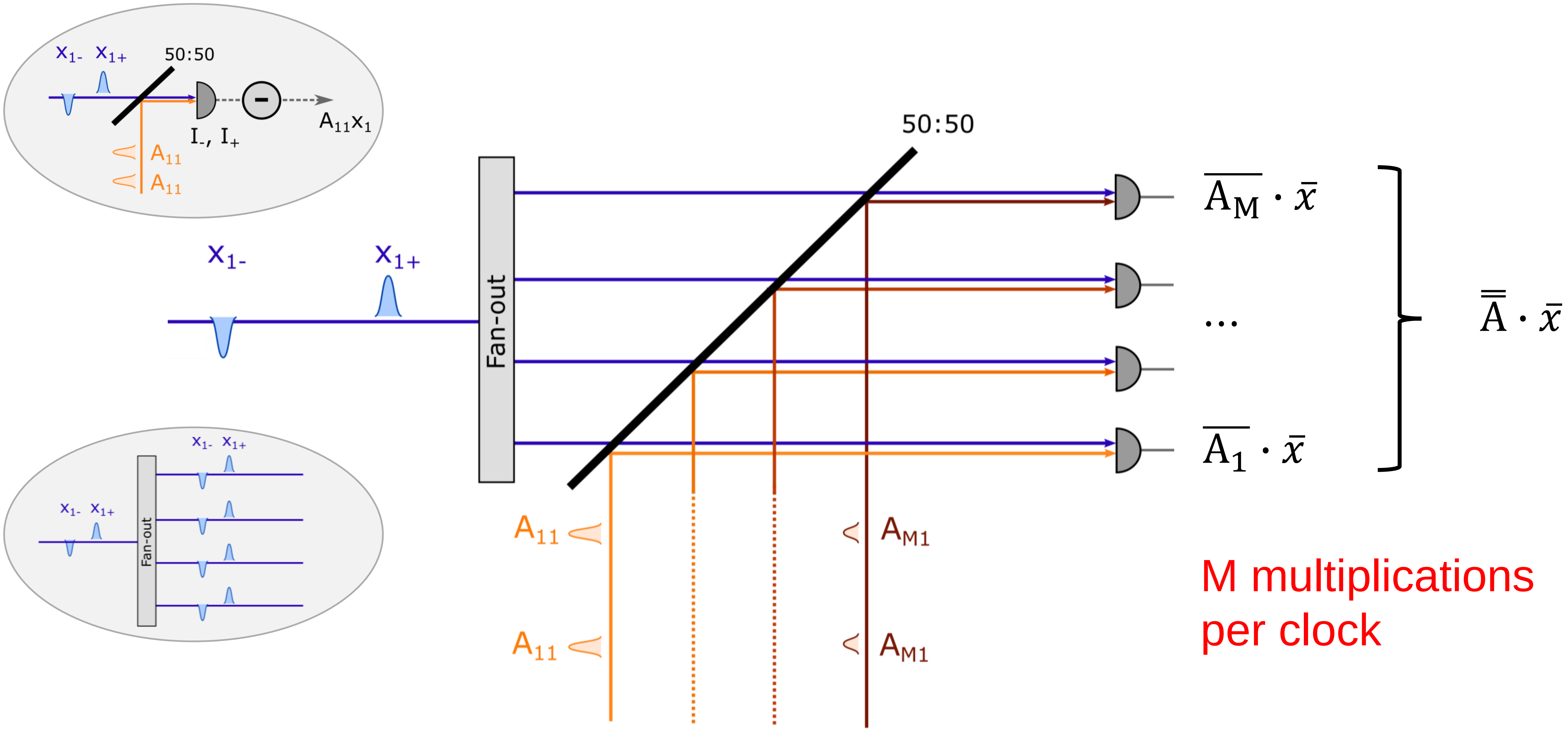
$$I_- = \left| \frac{1}{\sqrt{2}} A_{11} - \frac{i}{\sqrt{2}} x_1 \right|^2$$

$$\Big\downarrow \; I_+ \ominus I_-$$

$$\propto A_{11} x_1$$

[Hamerly, R., Bernstein, L., Sludds, A., Soljacic, M., Englund, D., *PRX* 2019]

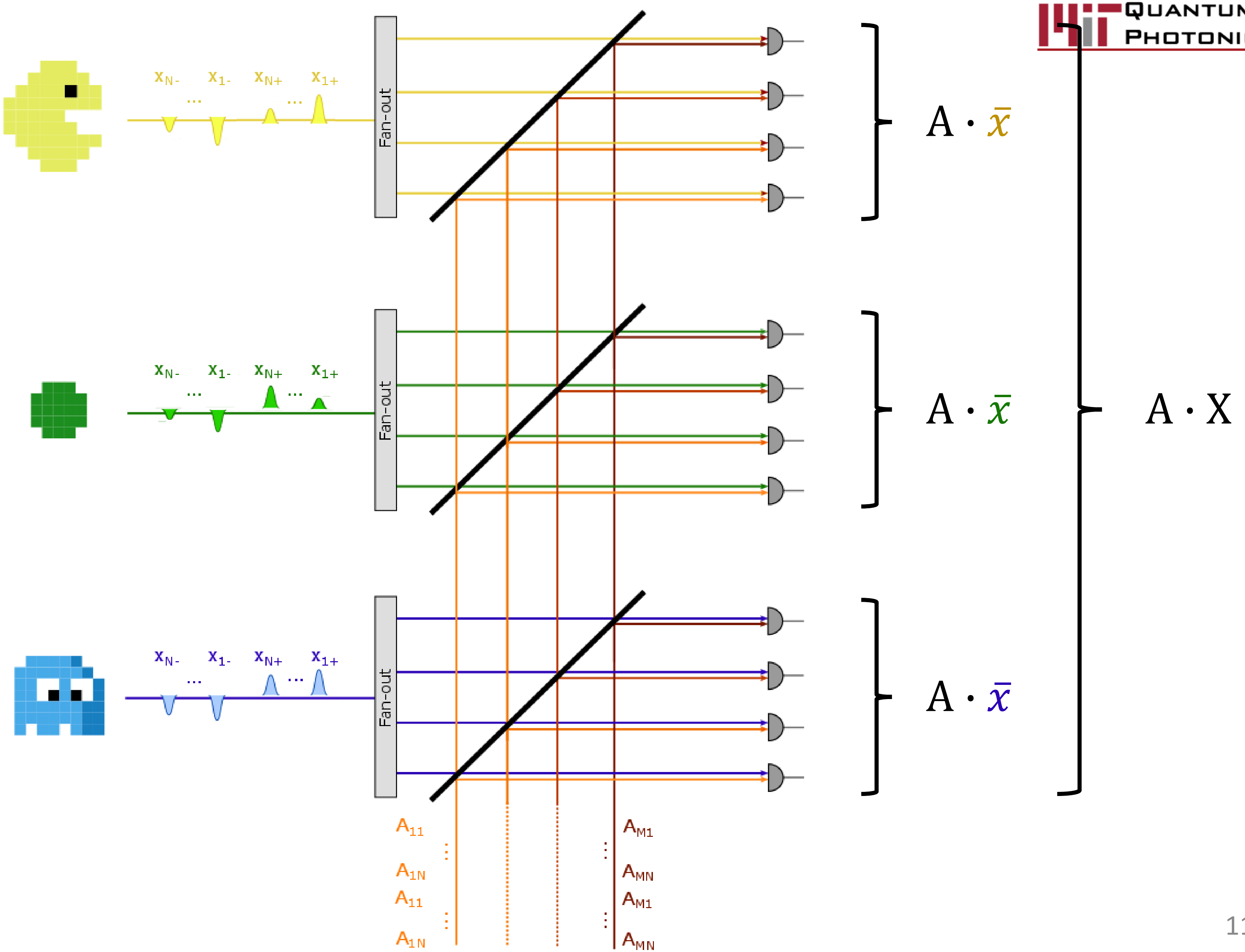
Optical multiplication + copying + summation



M multiplications per clock

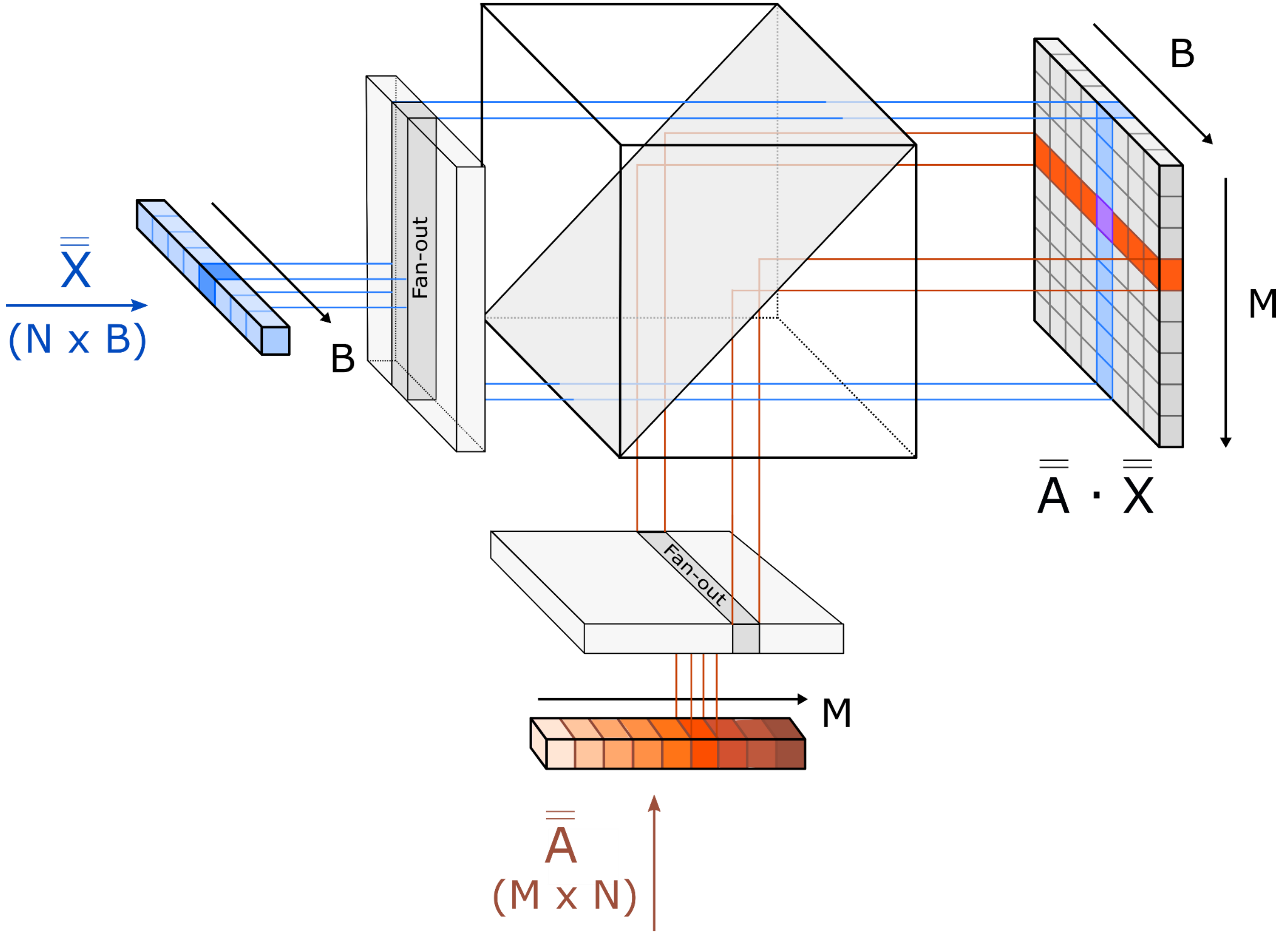
Optical routing to many clients

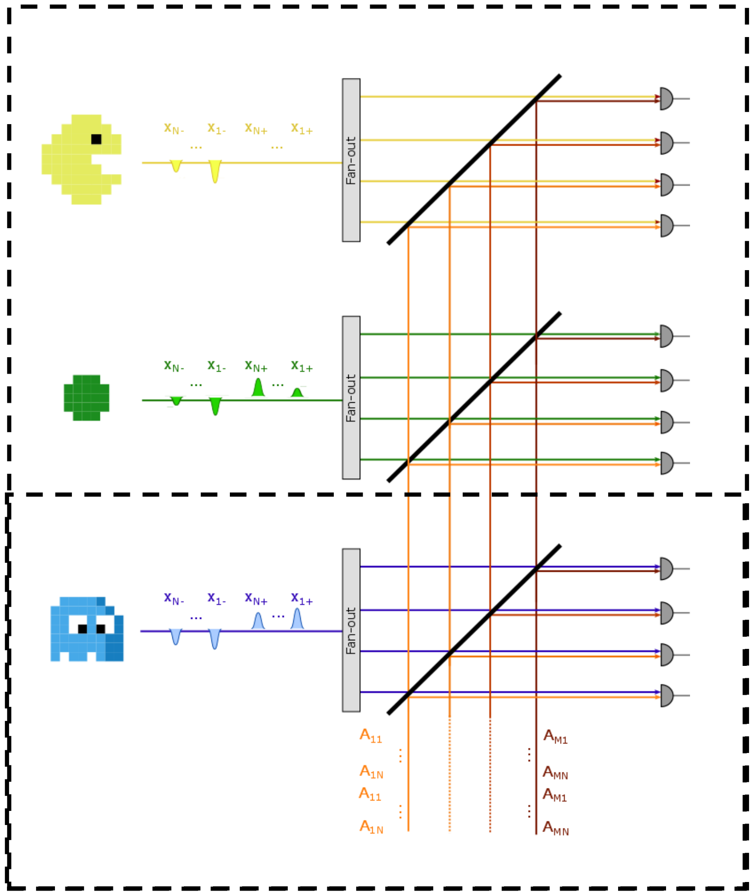
Amortized cost of weights



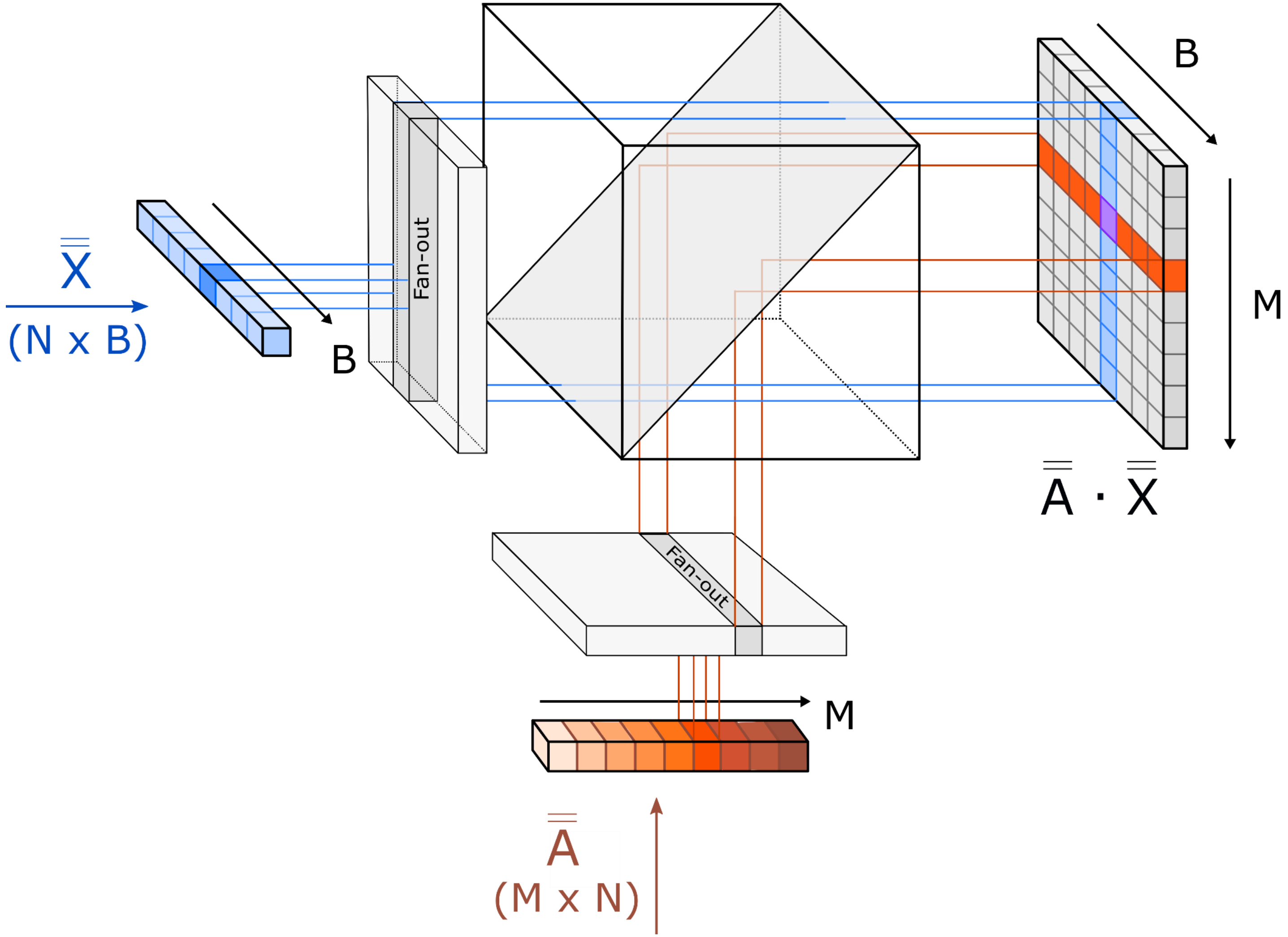
QUANTUM PHOTONICS

Optical Matrix-Matrix Multiplication





Optical Matrix-Matrix Multiplication



M * B multiplications per clock cycle

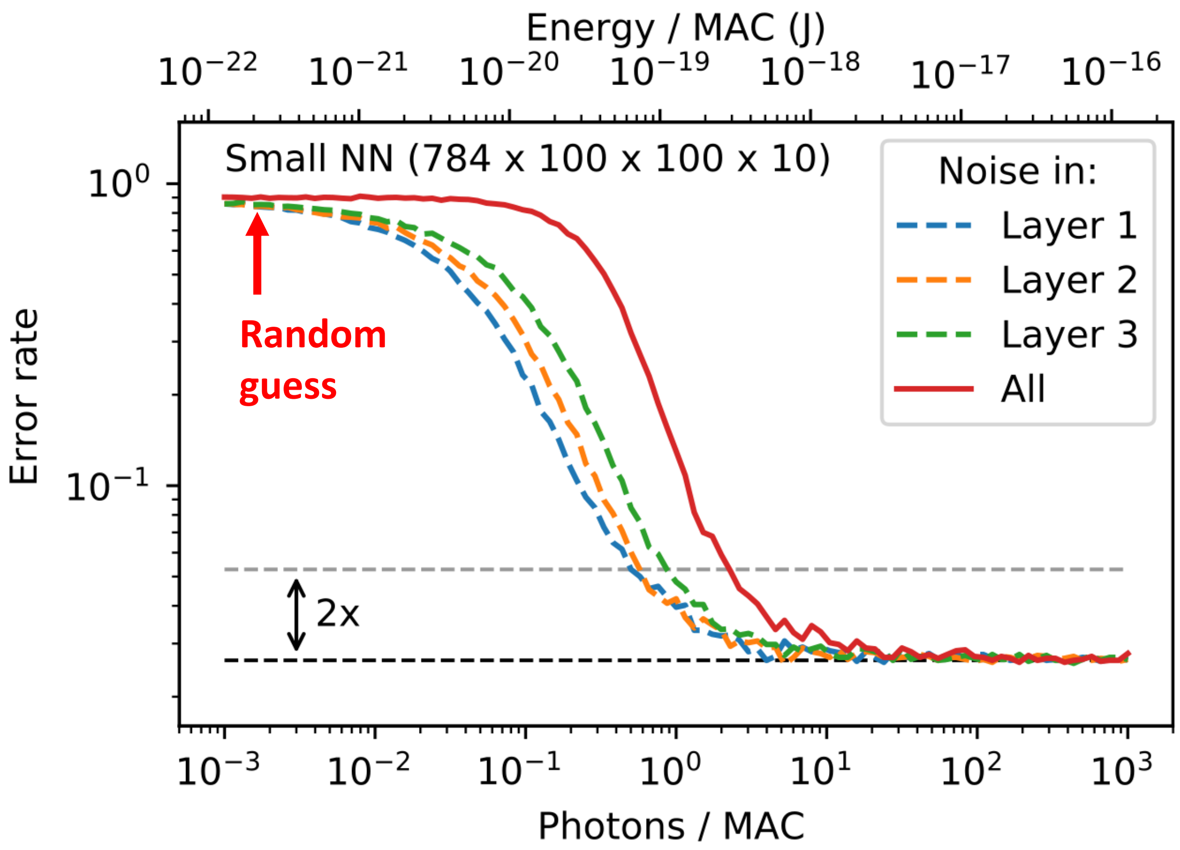
For N * M * B mults.:

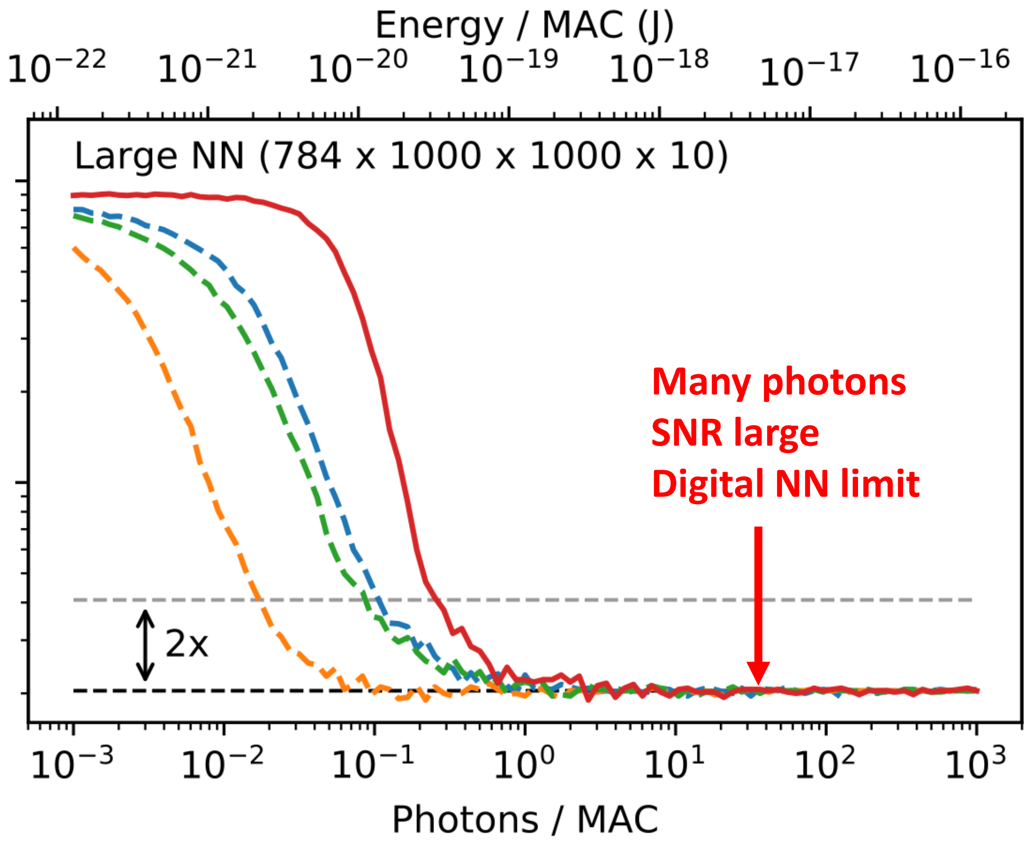
- N clock cycles
- B + M transmitters
- B * M receiver pixels

[Hamerly, R., Bernstein, L., Sludds, A., Soljacic, M., Englund, D., *PRX* 2019]

Theoretical Lower Energy Bound (Fundamental Limit)

- Fundamental limit of optics: error due to shot noise
- Simulations of classification error on MNIST dataset
- Lowest energy cut-off ≡ 2X canonical error rate





Theoretical Lower Energy Bound (Practical)

- Energy consumption: optical (shot limit)

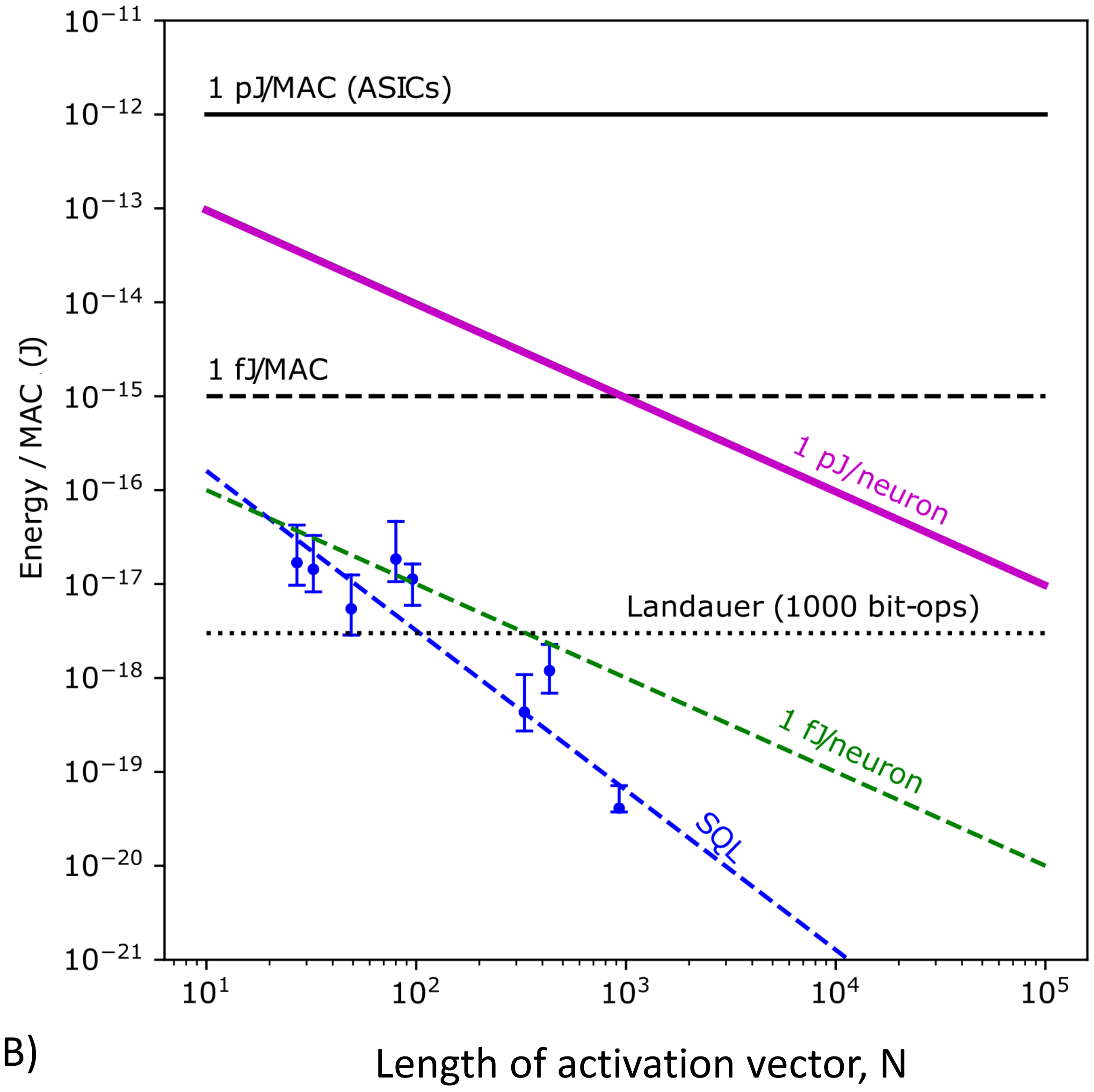
+ transmitter
+ receiver
+ electronics

Existing 1pJ/neuron

Emerging 1fJ/neuron

[DAB Miller, J. Lightwave Tech. 2017]

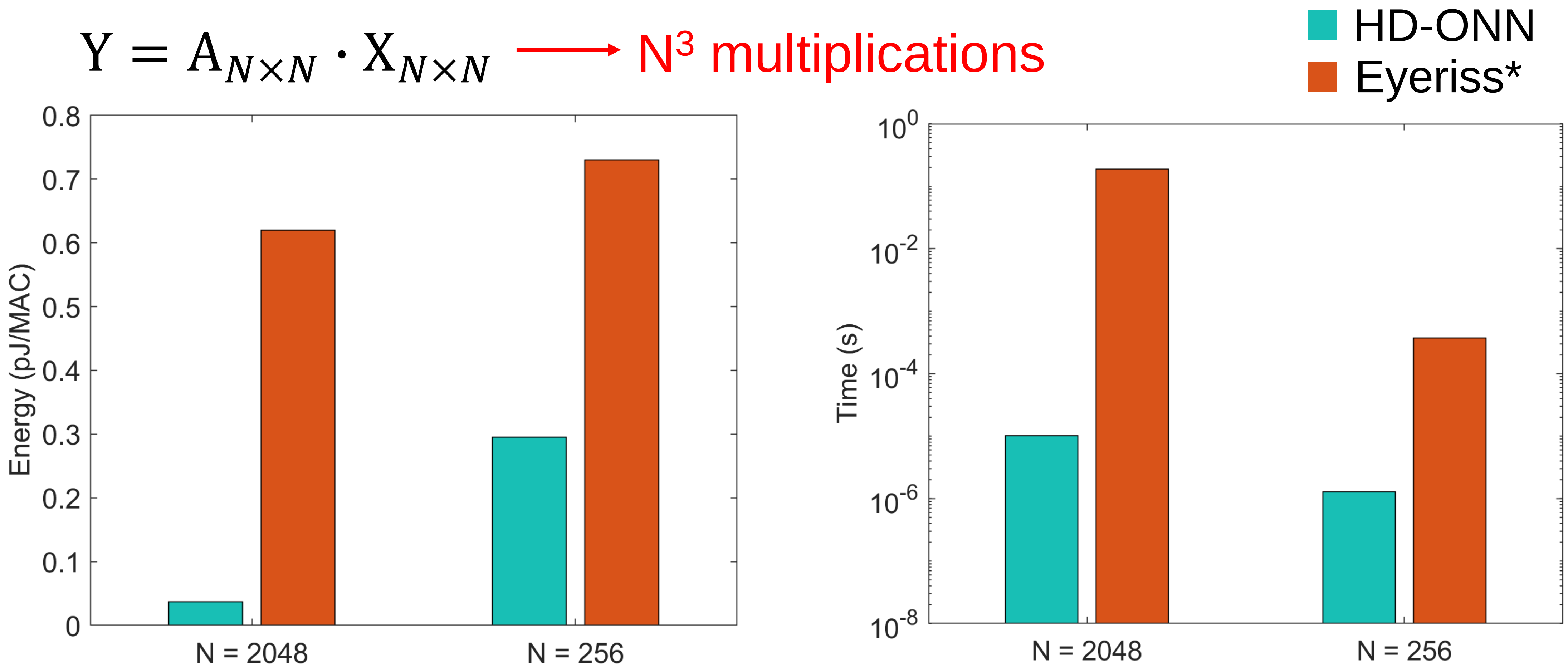
(assuming N = M = B)



Comparison of HD-ONN with Electronic NN Hardware

Energy and time required for matrix multiplication*

$$\mathrm{Y} = \mathrm{A}_{N\times N} \cdot \mathrm{X}_{N\times N} \longrightarrow N^3 \text{ multiplications}$$



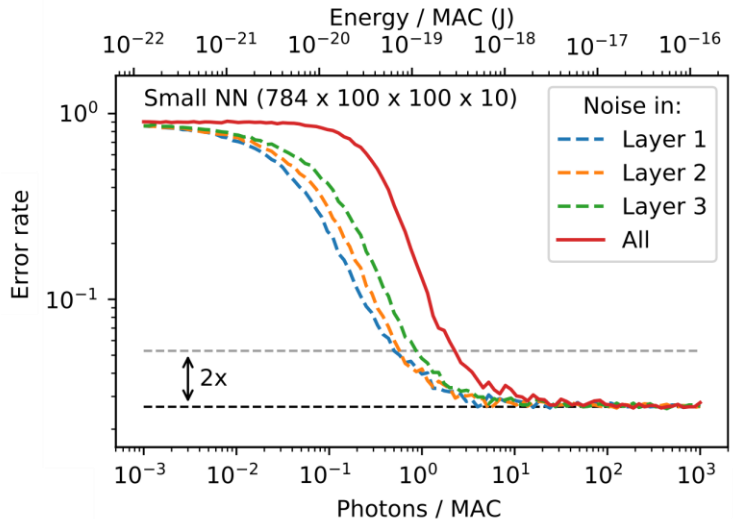
*includes memory access cost!

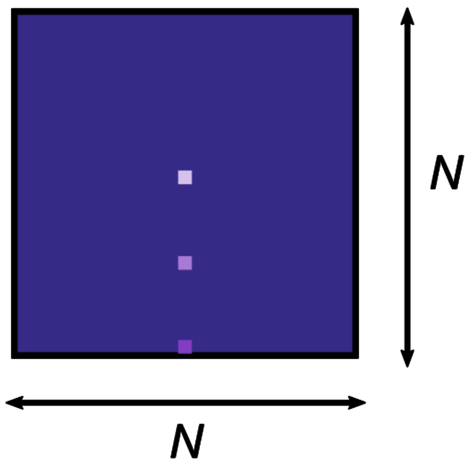
*[Sze, V. et al., *Proc. IEEE* 2017]
Hardware optimized for DNNs/CNNs with 256 processing elements

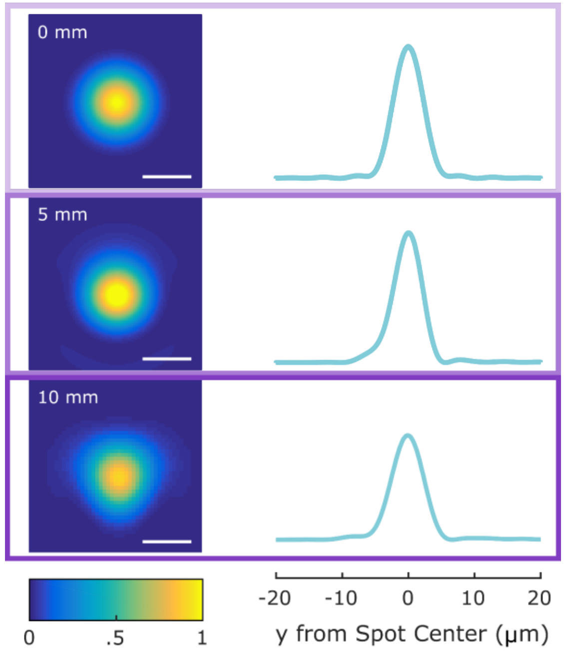
Other Considerations

Precision

- Shot noise
  - optical input power
- Optical crosstalk
  - packing of inputs and detectors





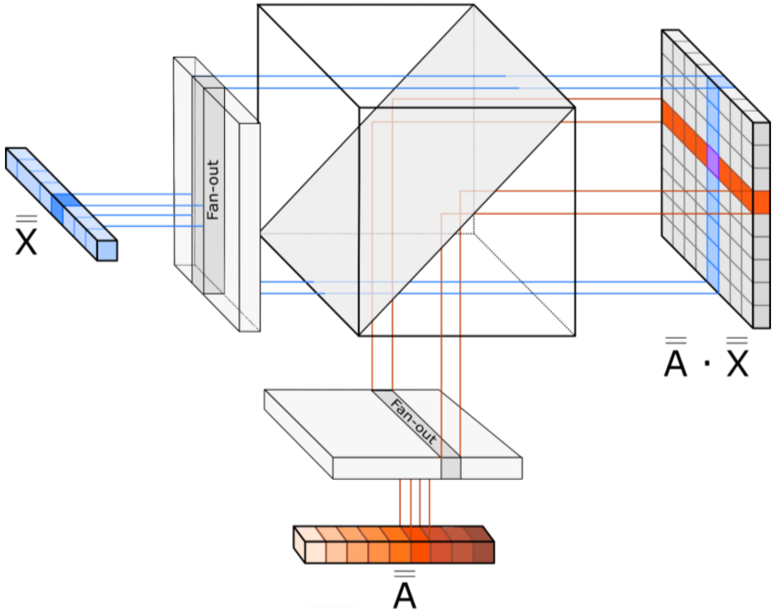


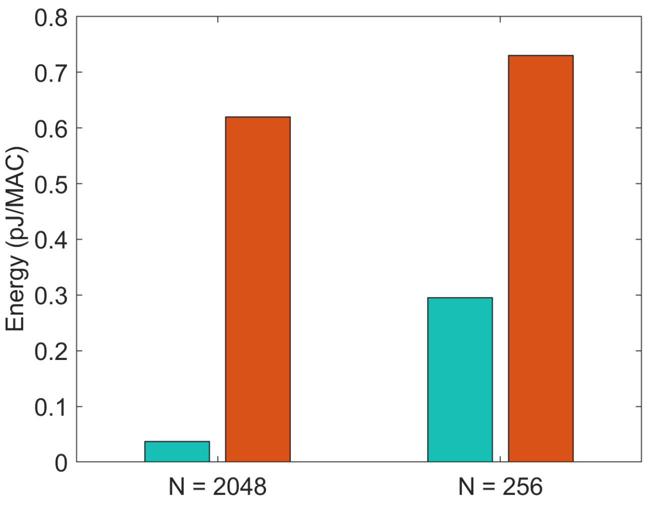
Area

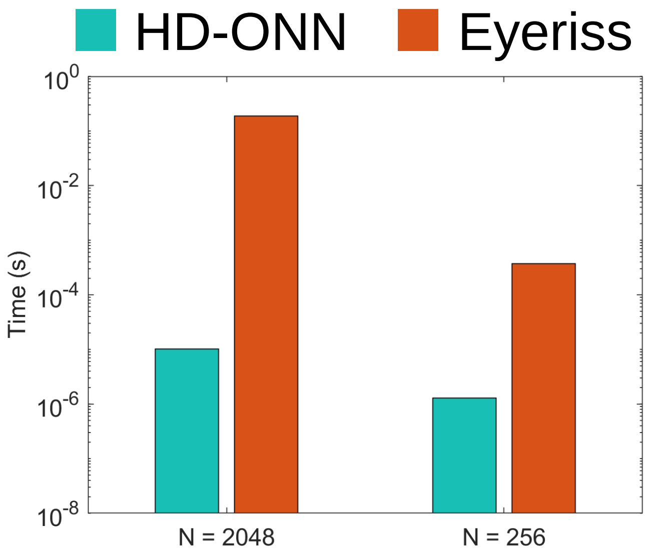
- Detector pixel i.e. neuron: ~(20 x 20) μm^2
- Transmitter: <50 μm
- Optics (lenses, beamsplitter)

Conclusions and Outlook

- Neural networks currently limited by energy consumption and speed
- Electronic accelerators reaching the limits of digital CMOS
- Optics provides new paradigm for neural networks and applications requiring large matrix multiplication
- Noise in analog regime (applications must be fault-tolerant)







Acknowledgements

MIT Quantum Photonics Group

- Prof. Dirk Englund
- Dr. Ryan Hamerly
- Alexander Sludds
- Lamia Ateshian

Collaborators

- Marin Soljačić
- Charles Roques-Carmes
- Joel Emer
- Vivienne Sze
- Yannan Wu
- Angshuman Parashar

Funding

NSERC CRSNG

OAK RIDGE INSTITUTE FOR SCIENCE AND EDUCATION

NVIDIA

AFRL THE AIR FORCE RESEARCH LABORATORY LEAD | DISCOVER | DEVELOP | DELIVER