LIGHTELLIGENCE

Accelerate AI with Light

LIGHTELLIGENCE

AI Computing with Integrated Photonics

Nov. 7th 2018

Artificial Neural Networks (ANN)

Breakthroughs in deep learning

- Computer vision
- Natural language processing (NLP)
- Game (Go, Atari)
- Autonomous driving
- Advertisement
- Drug discovery

Hungry for computing power:

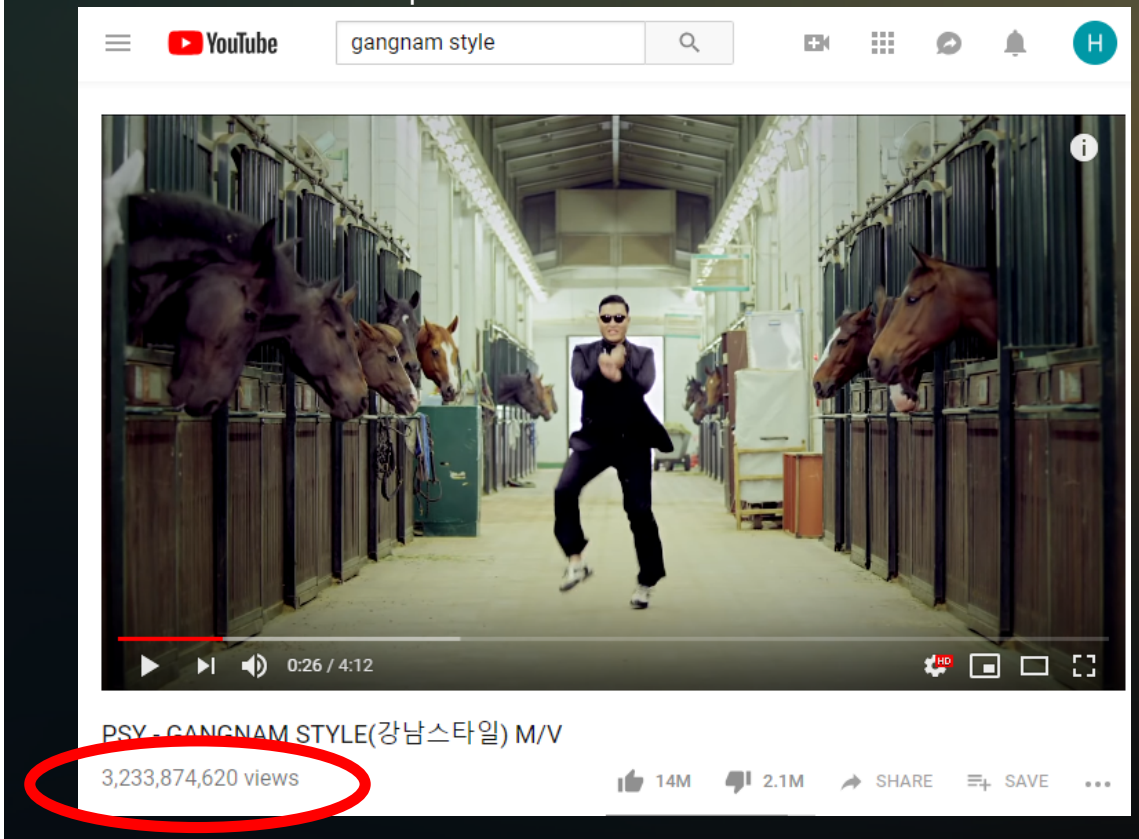
- More data
- Larger models

Google translate

LIGHTELLIGENCE

Data Growth

Popular videos in Youtube

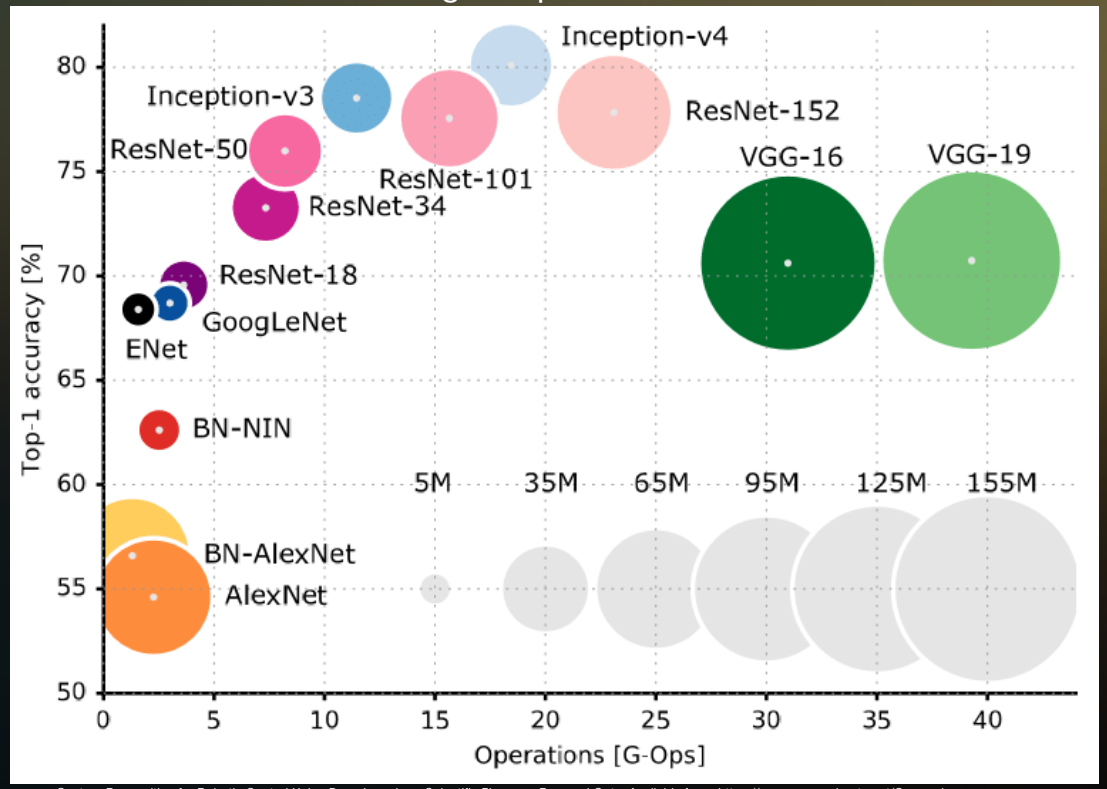


Projection of data growth

https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy

Price for Performance: Model Complexity

ImageNet performance

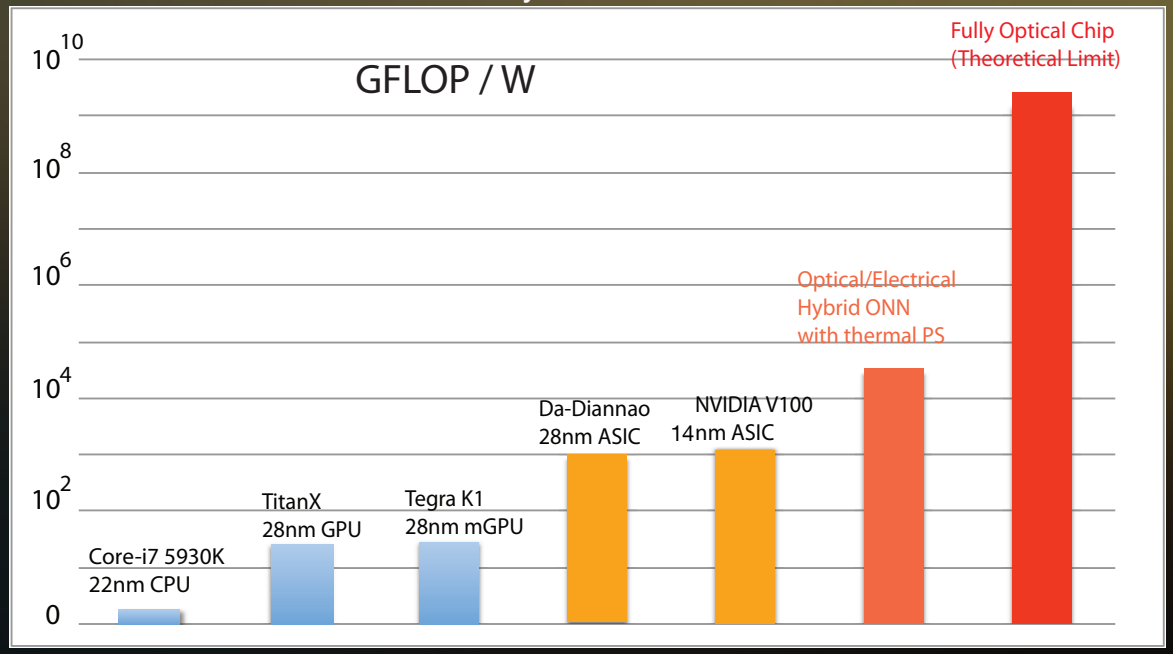


Gesture Recognition for Robotic Control Using Deep Learning - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Comparison-of-popular-CNN-architectures-The-vertical-axis-shows-top-1-accuracy-on_fig2_320084139 [accessed 31 Oct, 2018]

Need for Computation Power

- Larger model x more data = more computation
- Moore's law ends
- Specialized hardware for AI computing.
- Optical chips for next generation ASIC for AI

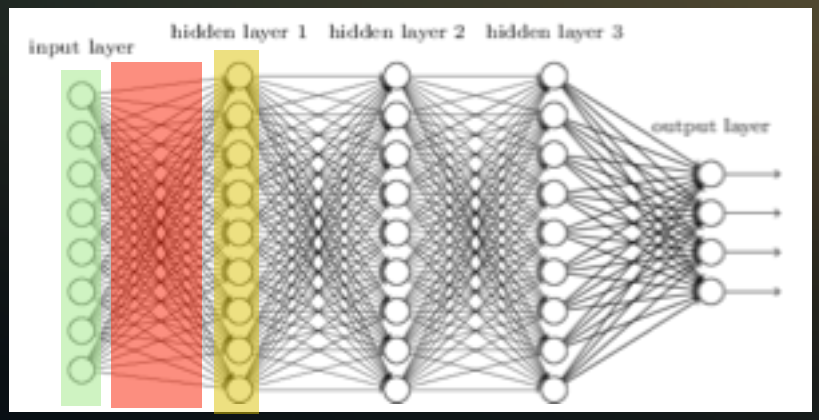
Power efficiency of AI hardware

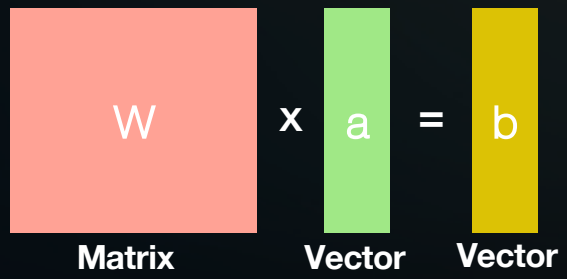


CPU → GPU → ASIC → Photonics

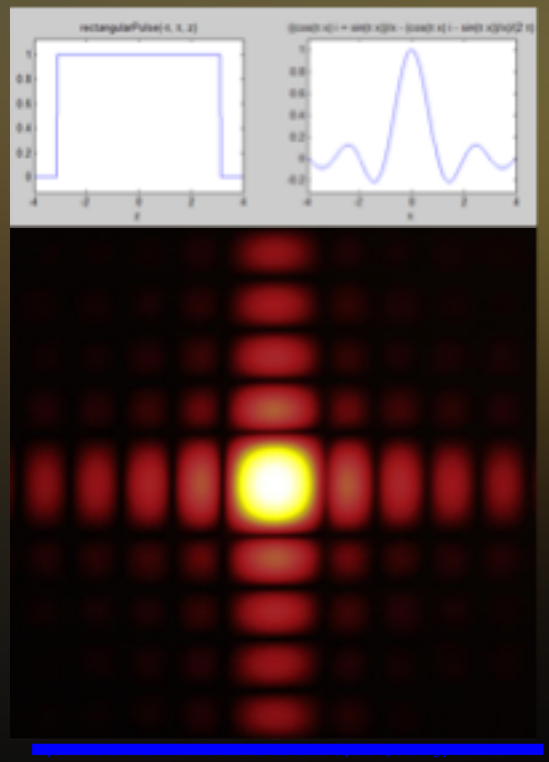
Optics for Linear Operation

- Most operations in neural network are linear
- Optical operation is linear in nature. Digital electronics are not.

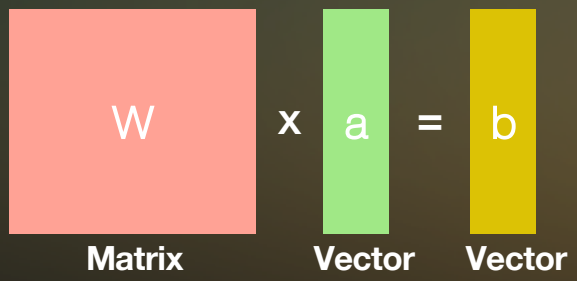


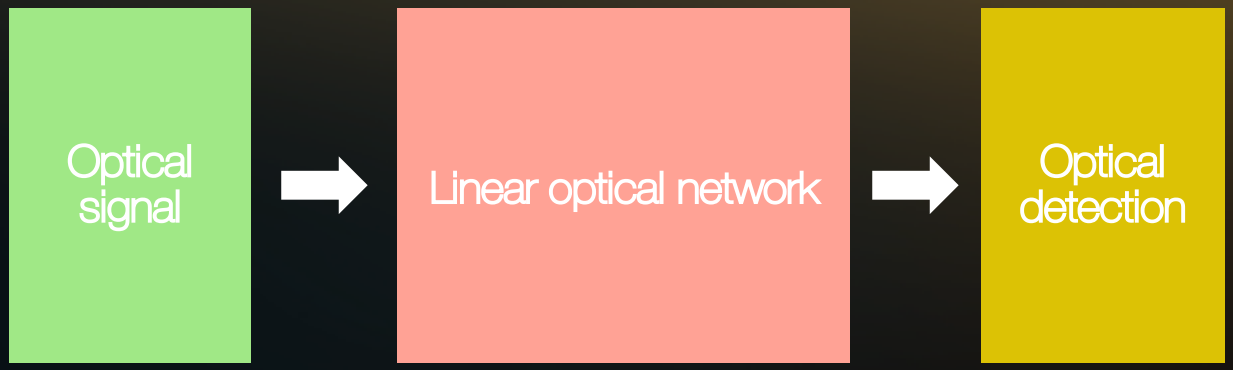


Diffraction of a square aperture:



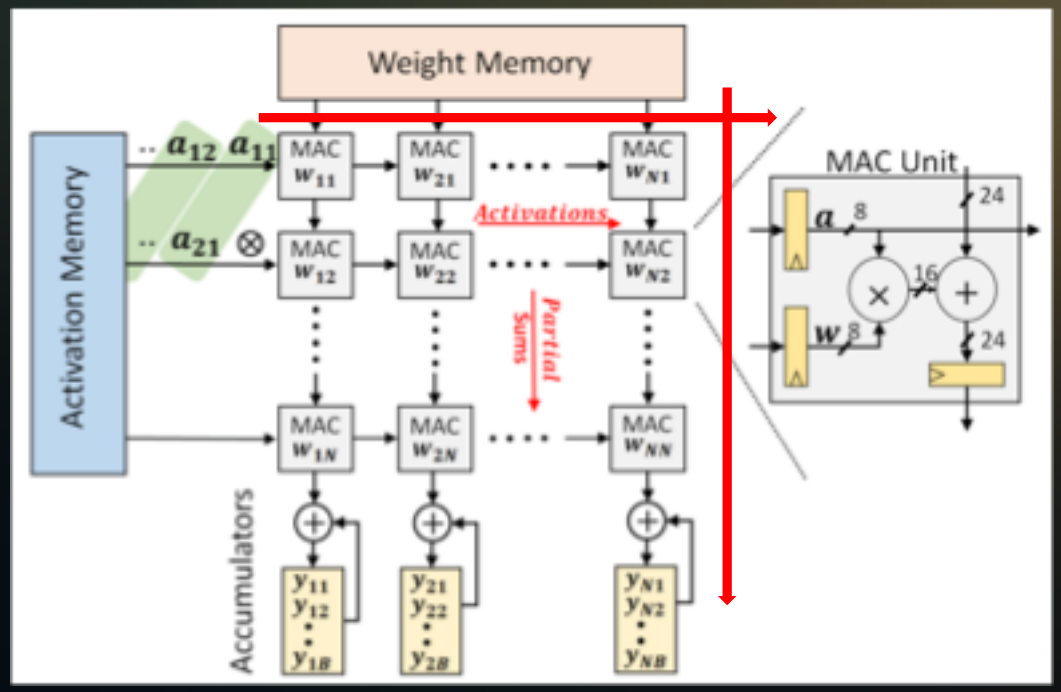
Optical Linear Operation





Optic Linear Operation: Latency

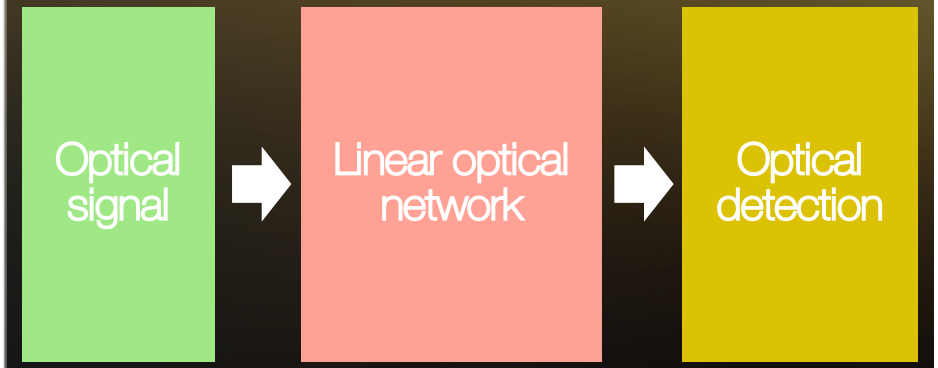
Systolic loop (TPU): a few hundred clock cycles



Zhang, Jeff, Kartheek Rangineni, Zahra Ghodsi, and Siddharth Garg. "Thundervolt: Enabling Aggressive Voltage Underscaling and Timing Error Resilience for Energy Efficient Deep Learning Accelerators." In *Proceedings of the 55th Annual Design Automation Conference on - DAC '18*, 1–6. San Francisco, California: ACM Press, 2018. [redacted].

Optics:

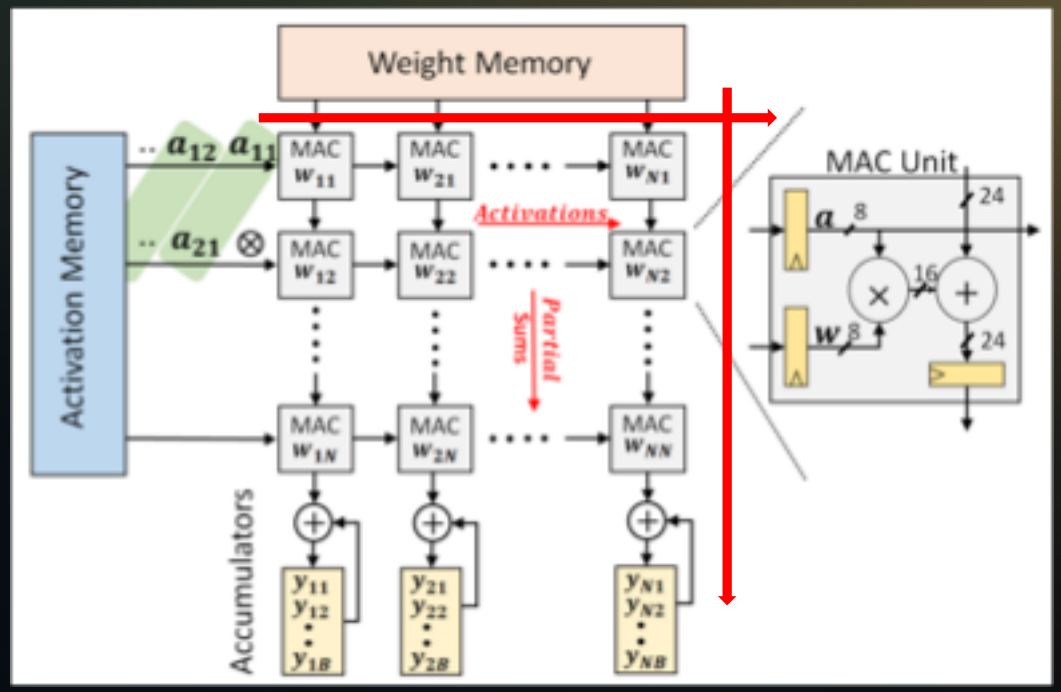
- Time of flight over chip
- Data modulation speed ~50GHz



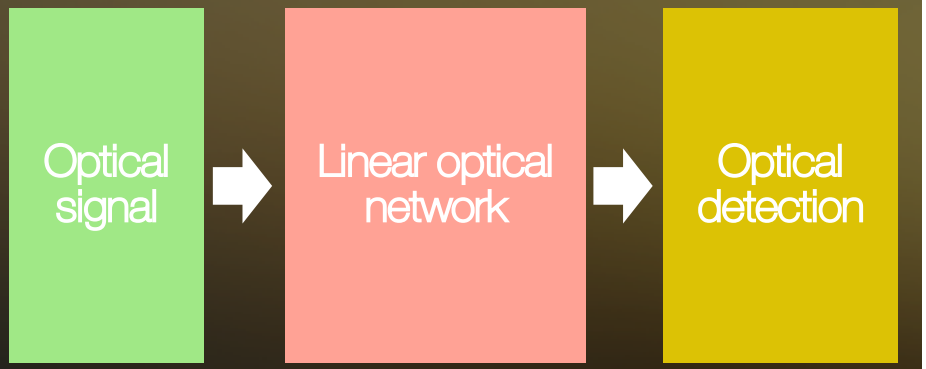
Optic Linear Operation: Energy Efficiency

For matrix of NxN

Systolic loop (TPU): O(N²)



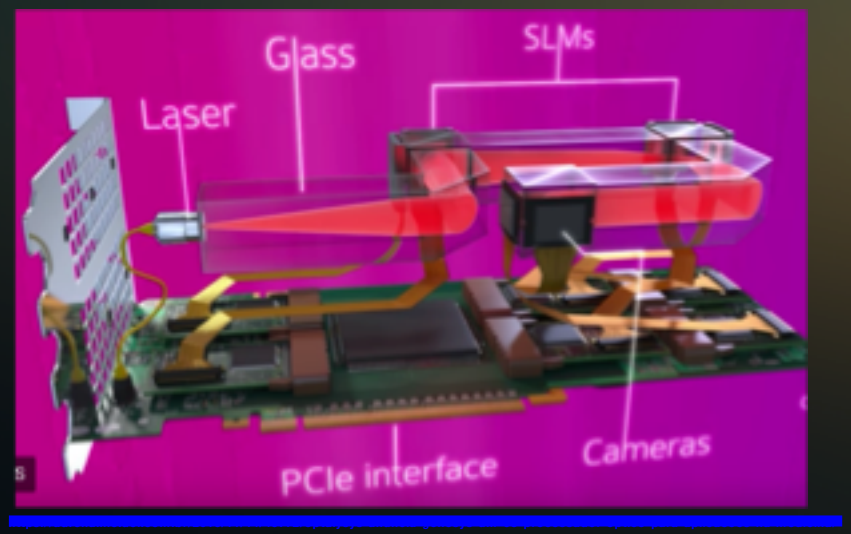
Optics: O(N)



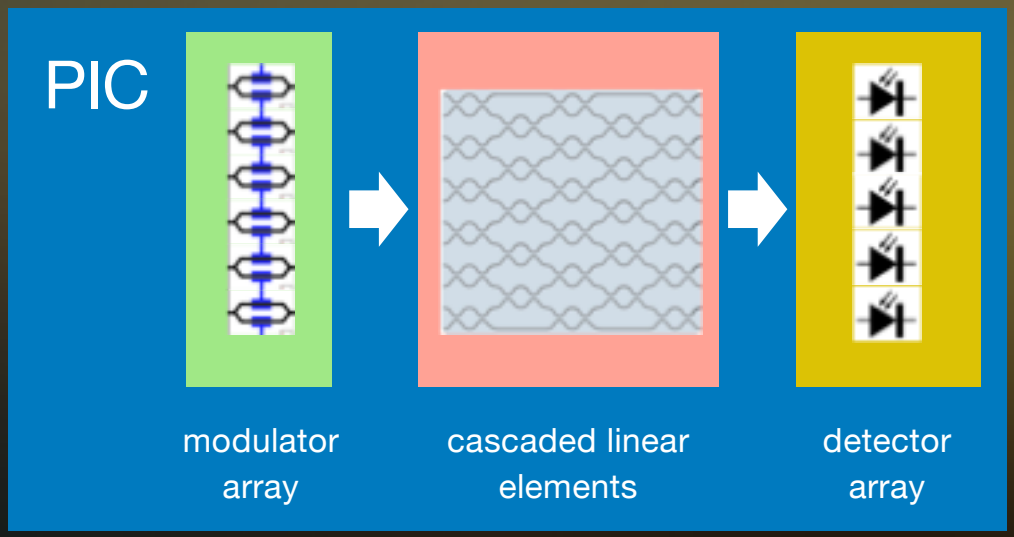
Zhang, Jeff, Kartheek Rangineni, Zahra Ghodsi, and Siddharth Garg. "Thundervolt: Enabling Aggressive Voltage Underscaling and Timing Error Resilience for Energy Efficient Deep Learning Accelerators." In *Proceedings of the 55th Annual Design Automation Conference on - DAC '18*, 1–6. San Francisco, California: ACM Press, 2018. [redacted].

Physical Implementation

Free space optics



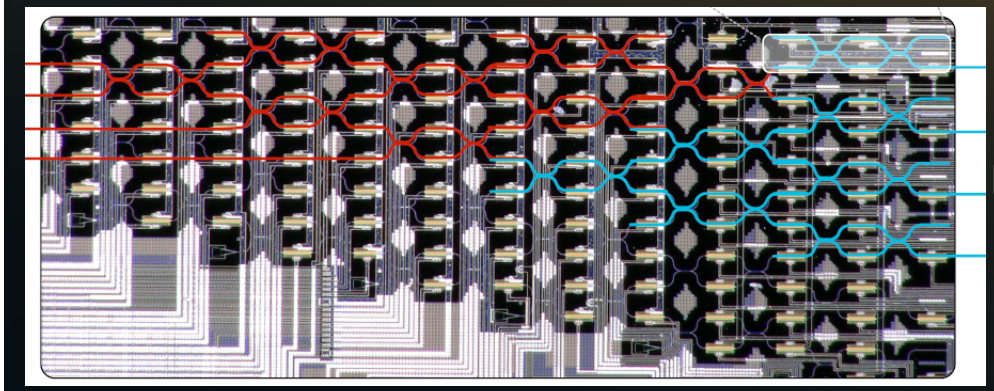
Photonic integrated circuit (PIC)



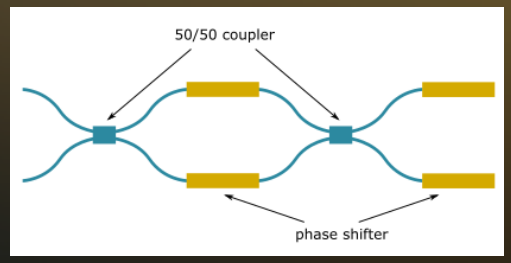
- Leverage mature semiconductor industry
- Robust against environment
- High data rate

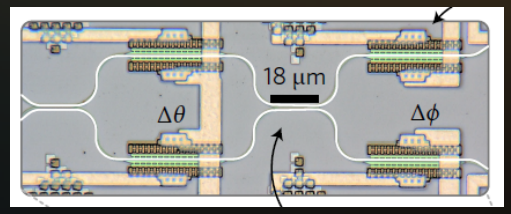
Cascaded MZI as Linear Network

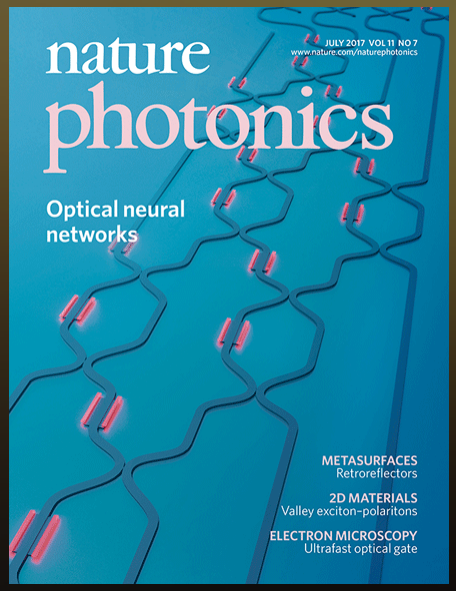
Optical signal → Linear optical network → Optical detection



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







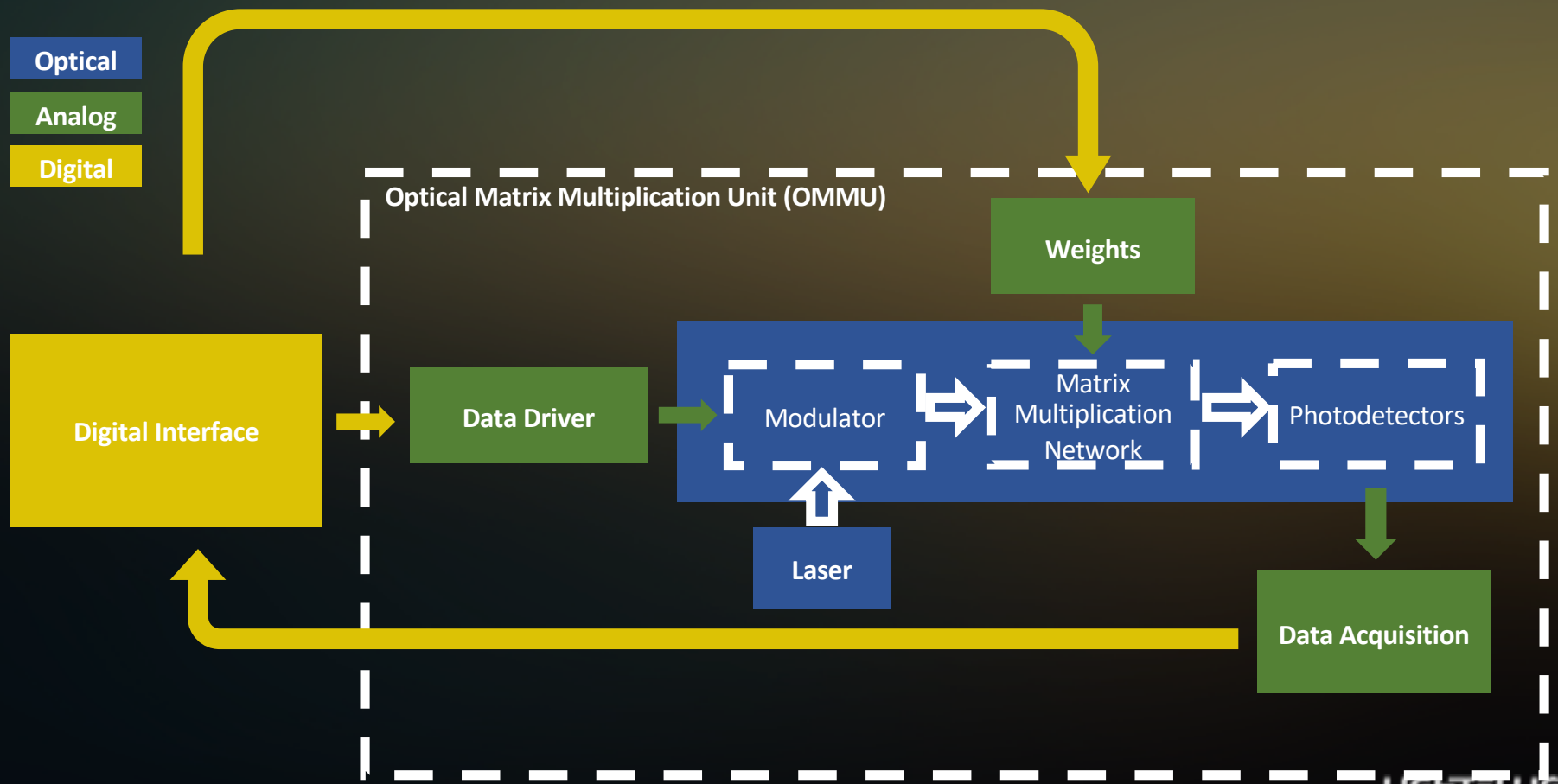
Shen, Yichen, Nicholas C. Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, et al. "Deep Learning with Coherent Nanophotonic Circuits." *Nat Photon* 11 (print 2017): 441–46. [redacted].

LIGHTELLIGENCE

Our Focus Areas

- System architecture
- Algorithm
- Analog electronics
- Fabrication non-uniformity
- System in package

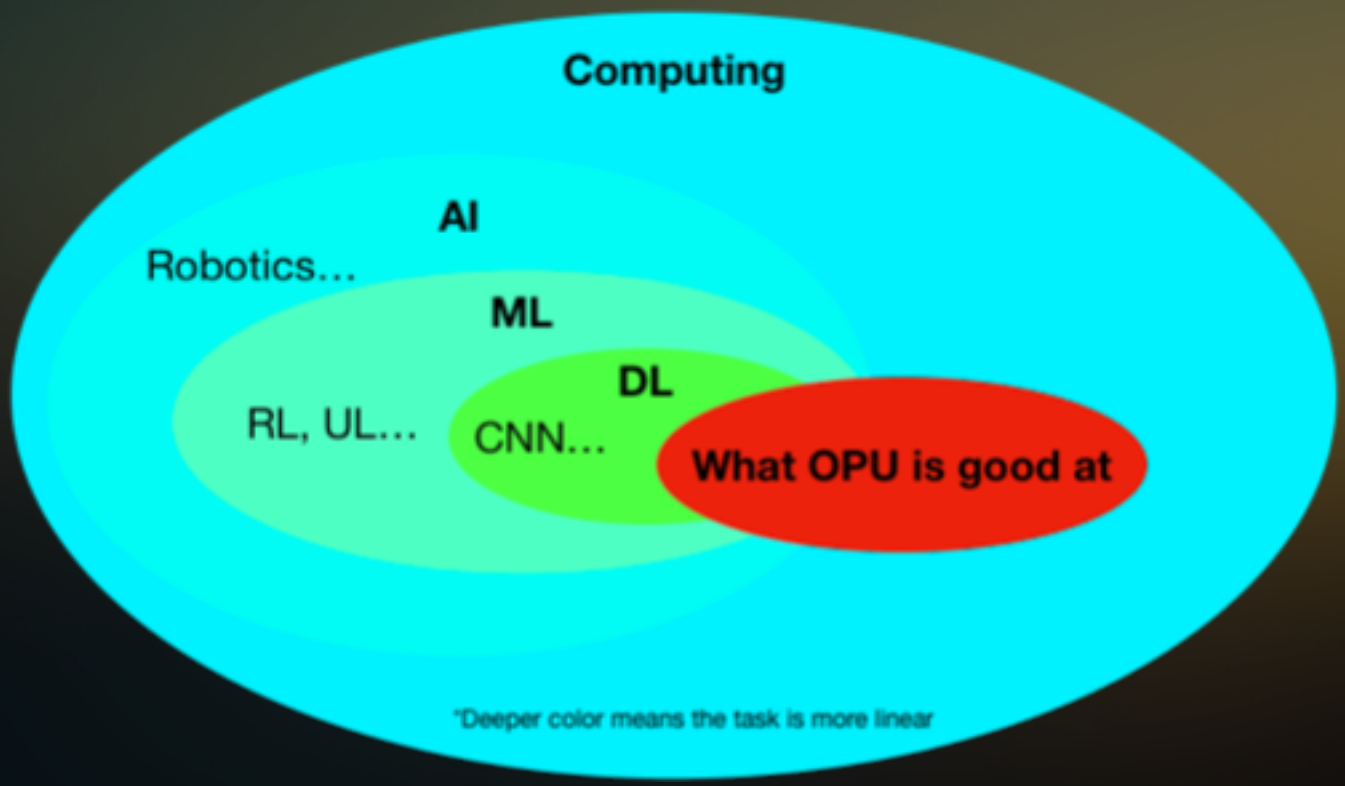
PoC System for Optical Computing



System Architecture Challenges

- High demand on data communication efficiency
- Data fetching and transport latency
- Advanced architectures and circuit techniques will be needed

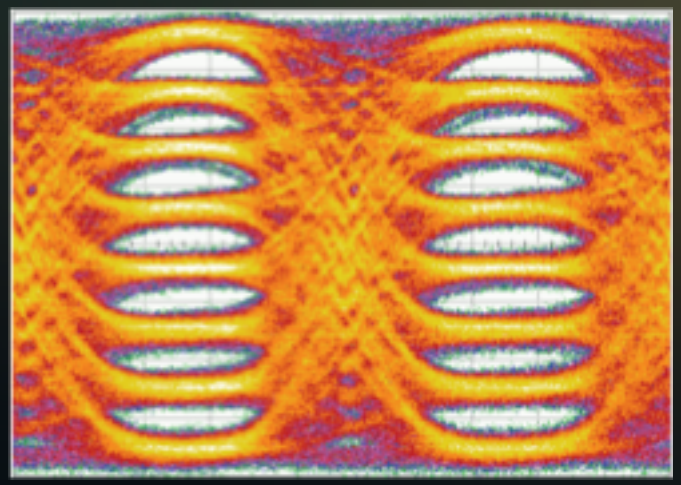
Algorithm Considerations



- OPU (Optical Processing Unit) is best for highly linear and low memory cost computing tasks.
- There is a sweet point in Deep learning for Optical Computing.
- It's also important to jump out of the scope of AI for more opportunities.

ANN with Low Bit Depth

Optic PAM8 eye diagram



Karlsson, M., E. Agrell, K. Szczerba, P. Andrekson, and A. Larsson. "35.2 Gbps 8-PAM Transmission Over 100 m of MMF Using an 850 Nm VCSEL." In *39th European Conference and Exhibition on Optical Communication (ECOC 2013)*, 744–46. London, UK: Institution of Engineering and Technology, 2013.

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)

LIGHTELLIGENCE

Analog Electronic Considerations

- O/E E/O conversion overhead
- Maintain signal-noise ratio at high bandwidth
- RF crosstalk

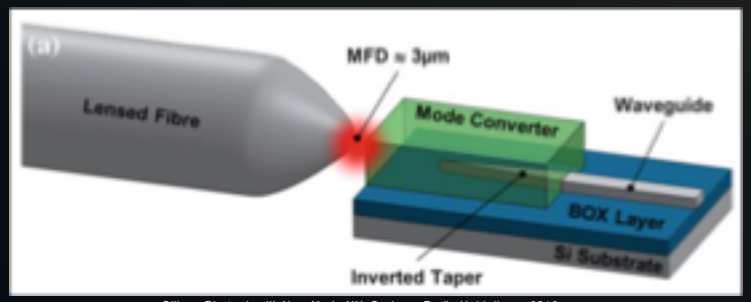
Fabrication Non-uniformity

- Consistency between multiple optical channels
- Source of fabrication variation:
  - SOI silicon thickness
  - Lithography accuracy
  - Surface/side roughness
  - Doping profile repeatability

System in Package

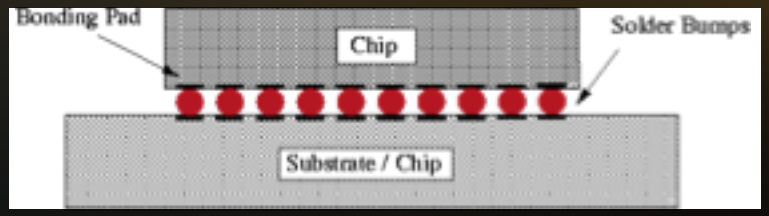
- Integration of optical and electric packaging
- High pin count
- High speed RF components

Optic packaging



Silicon Photonics III. New York, NY: Springer Berlin Heidelberg, 2016.

Electric packaging



Latency – Focus Areas

In addition to latency from optical processing:

- Data access
- E/O O/E conversion
- Weights shuffle
- Nonlinear operation

Energy Efficiency – Focus Areas

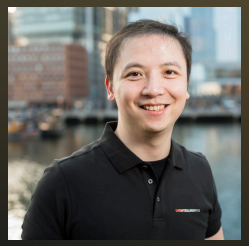
In addition to power consumption from optical processing:

- Data access and routing
- E/O O/E conversion (A/D, D/A conversion)
- Maintain weights
- Laser

Our Company

- Lightelligence is incorporated in late 2017
- Team consists of MIT PhD, professors, industry veterans
- $10M A round funding

Dr. Yichen Shen
Co-Founder, CEO



Dr. Paul Xie
Co-Founder, VP of Product

Dr. Huaiyu Meng
Co-Founder, VP of Photonics

Li Jing
Co-Founder
Chief Algorithm Architect

Maurice Steinman
VP of Engineering

Dr. Marin Soljacic
Co-Founder, Board Member

Dr. John Joannopolulos
Co-Founder, Advisor

LIGHTELLIGENCE

THANKS