NEUROMORPHIC SILICON PHOTONICS AND APPLICATIONS

Bhavin J. Shastri & Paul R. Prucnal

Queen's University

VECTOR INSTITUTE

PRINCETON UNIVERSITY

Collaborators

Paul Prucnal
Princeton

Lukas Chrostowski
UBC

Sudip Shekhar
UBC

Volker Sorger
GWU

+ many others and co-authors

Fabrication Support

CMC MICROSYSTEMS

amf ADVANCED MICRO FOUNDRY

SiEPIC

ANT APPLIED NANOTOOLS INC.

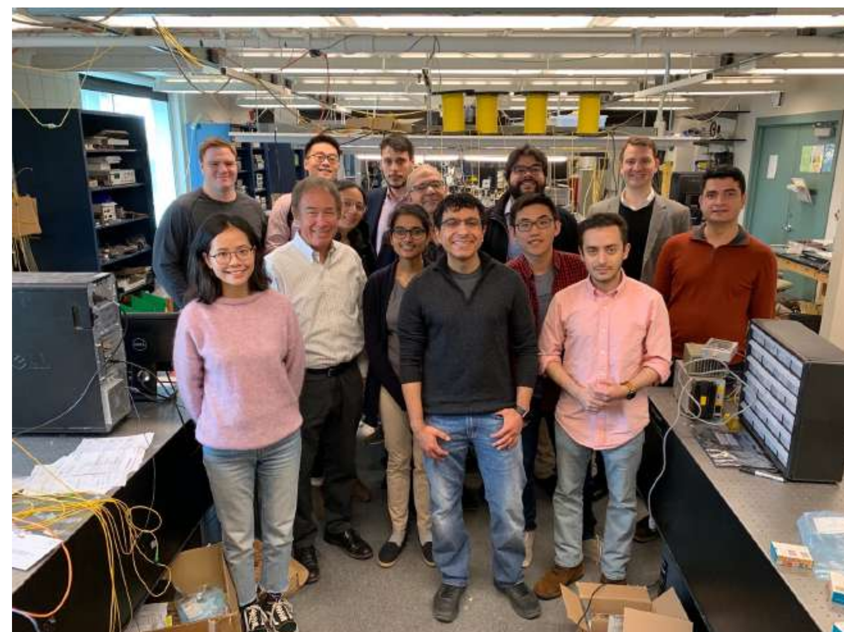
Students and Postdocs

Shastri Lab (Queen's)

Bicky Marquez (Postdoc)
Matthew Filipovich (Grad. Student)
Zhimu Guo (Grad. Student)
Hugh Grant (Grad. Student)
Jagmeet Singh (Grad. Student)
Marcus Tamura (Grad. Student)
Maryam Moridsadat (Grad. Student)
Ahmed Khaled (Grad. Student)
Karanpreet Singh (Grad. Student)

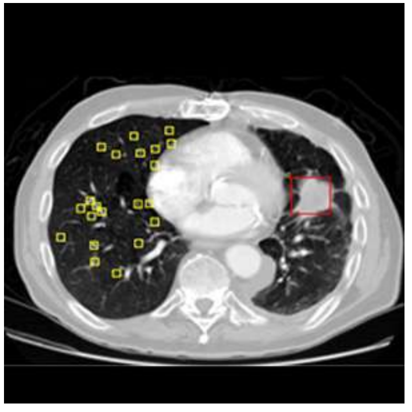
Lightwave Prucnal Lab (Princeton)

Paul R. Prucnal (PI)
Chaoran Huang (Postdoc)
Thomas Ferreira de Lima (Grad. Student)
Hsuan-Tung Peng (Grad. Student)
Eric Blow (Grad. Student)
Aashu Jha (Grad. Student)
Simon Bilodeau (Grad. Student)
Weipang Zhang (Grad. Student)
Josh Lederman (Grad. Student)
+ former students



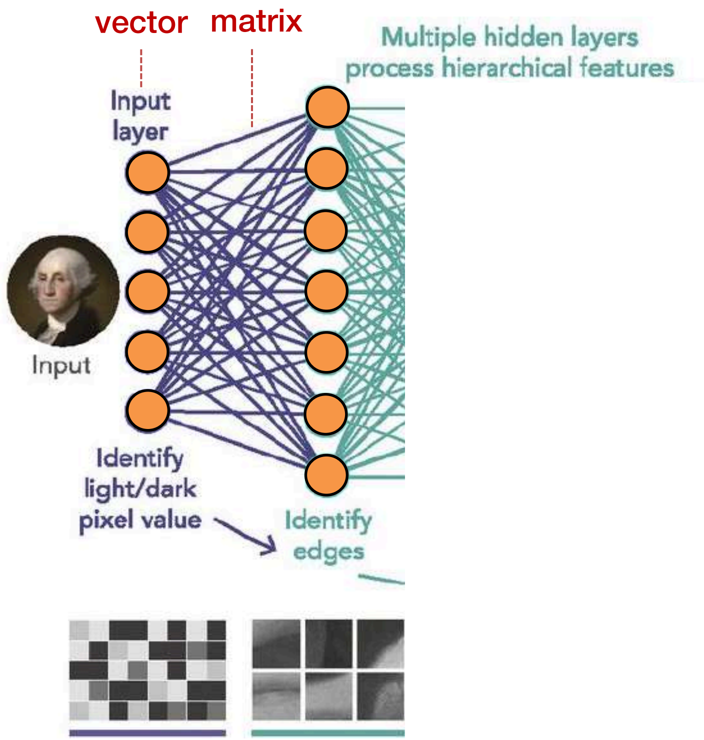
Rise of Artificial Intelligence

- Natural language processing; Siri, Cortana, Alexa, Google
- Game playing (Go) and Chess; beating World Champions
- Youtube, Netflix, Google Maps, Amazon
- Face recognition
- Autonomous vehicles
- Medicine and healthcare
- Control and optimization



LeCun, Bengio, Hinton *Nature* **521** (2015)

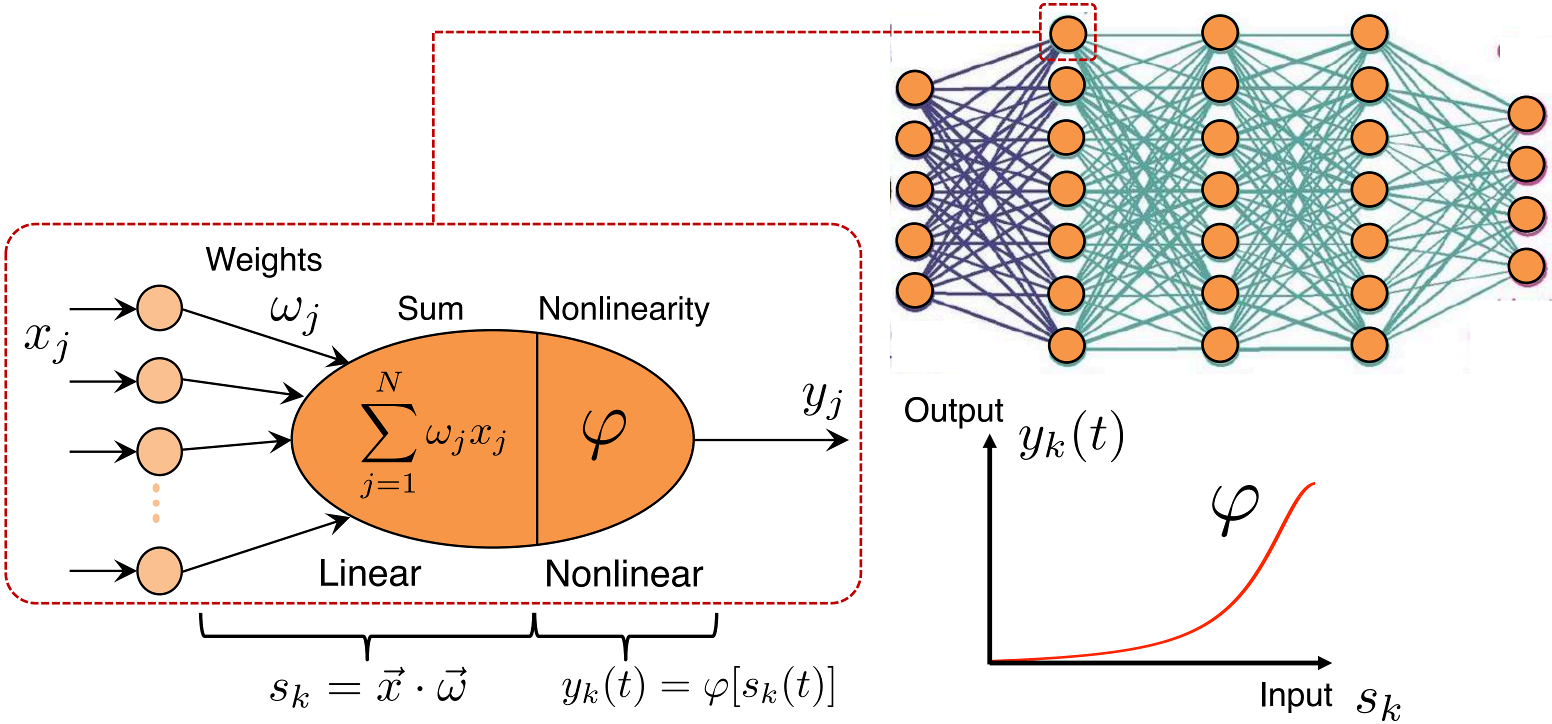
Deep Neural Networks Enable AI



- Neural networks have many layers (with many neurons)
- Layers are connected with weights
- Each layer needs to be trained (e.g. backpropagation)
- Dense vector-matrix multiplications
- Data movement takes resources

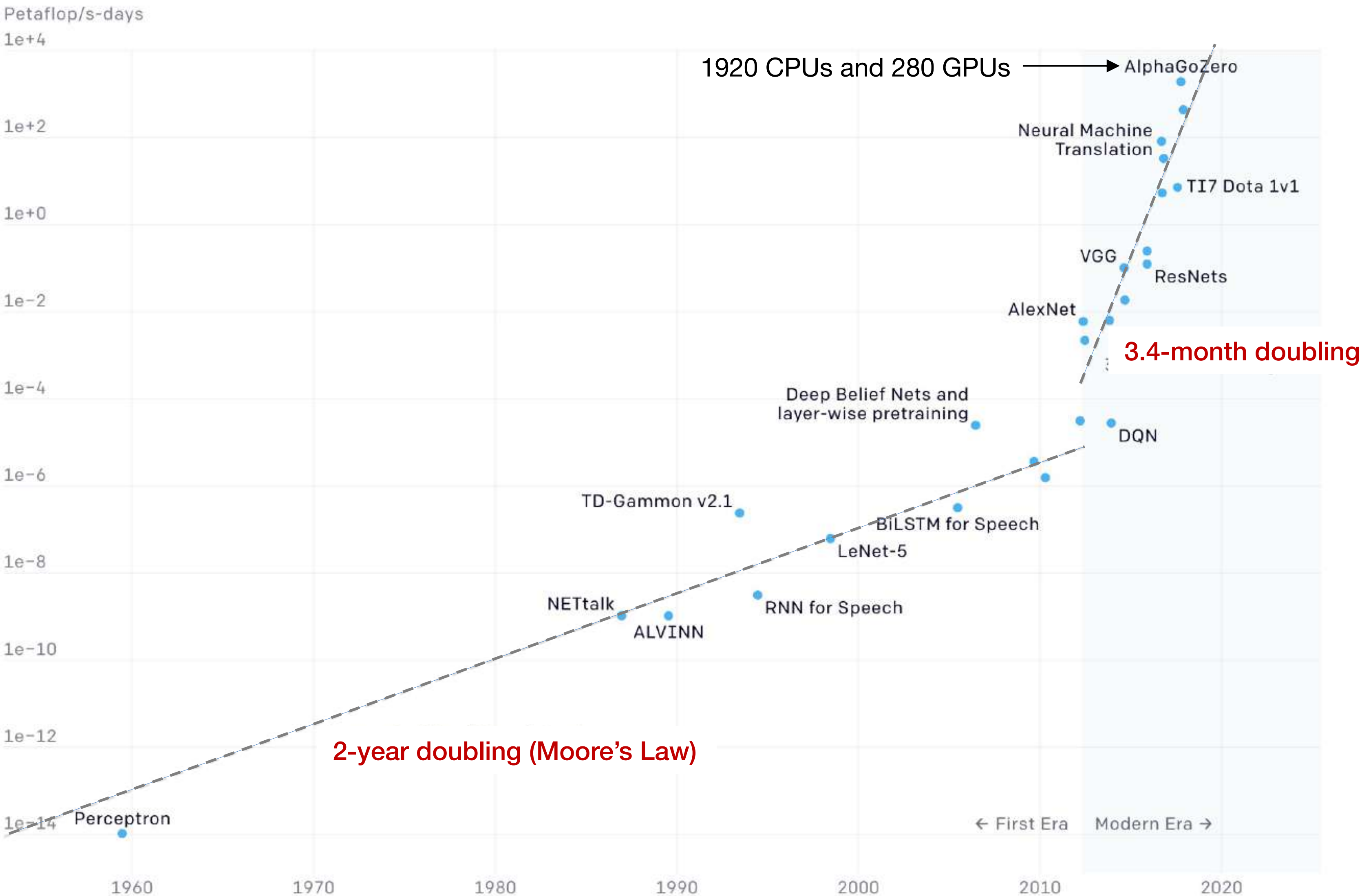
LeCun, Bengio, Hinton *Nature* **521** (2015)
Waldrop *PNAS* **116** (2019)

Neuron Model and Weighted Sum



- **Neuron operation:** 1) receive multiple inputs; 2) weight each by coefficient; 3) sum them; 4) perform a nonlinear thresholding
- Linear: 1-3 is a weighted sum or **multiply & accumulate** (MAC)
- Nonlinear: state variable s is thresholded by a nonlinearity ϕ

Two Distinct Eras of Compute Usage in AI Systems



https://blog.openai.com/ai-and-compute/

Neuromorphic Hardware for AI

	Function	Electronics	Photonics**
Computing	Nonlinearity	Easy (transistors)	Hard
	Memory	Easy (DRAM, SRAM)	Hard
	Gain	Easy	Easy (but off-chip)
Communication	Communication	~CV^2 Energy Cost	Free (waveguide)
	Fan-out	~CV2 Energy Cost	Free (beamsplitter)
Linear Algebra	MAC & MVM	Hard*	Easy (EO Components)
System	Domain Crossings	No	Yes*** (O-E-O)

Notes

* For digital electronics

** State-of-art, e.g. Review in Shastri et al. *Nature Phot*. (2021)
Shen et al. *Nature Phot*. (2017)

*** Tait et al. *Sci. Rep.* (2017), Tait et al. *Phys. Rev. Appl.* (2018)

O = optical, E = Electrical

Credit: Ryan Hamerly (MIT) and Volker Sorger (GWU)

REVIEW ARTICLE | FOCUS
https://doi.org/10.1038/s41566-020-00754-y
nature photonics

Check for updates

Photonics for artificial intelligence and neuromorphic computing

Bhavin J. Shastri[1,2,7] ✉, Alexander N. Tait[2,3,7] ✉, T. Ferreira de Lima[2], Wolfram H. P. Pernice[4], Harish Bhaskaran[5], C. D. Wright[6] and Paul R. Prucnal[2]

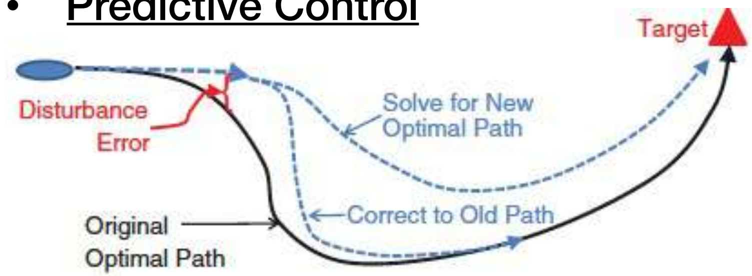
Research in photonic computing has flourished due to the proliferation of optoelectronic components on photonic integration platforms. Photonic integrated circuits have enabled ultrafast artificial neural networks, providing a framework for a new class of information processing machines. Algorithms running on such hardware have the potential to address the growing demand for machine learning and artificial intelligence in areas such as medical diagnosis, telecommunications, and high-performance and scientific computing. In parallel, the development of neuromorphic electronics has highlighted challenges in that domain, particularly related to processor latency. Neuromorphic photonics offers sub-nanosecond latencies, providing a complementary opportunity to extend the domain of artificial intelligence. Here, we review recent advances in integrated photonic neuromorphic systems, discuss current and future challenges, and outline the advances in science and technology needed to meet those challenges.

Shastri, Tait et al. *Nature Photonics* **15** (2021)

Applications for Photonic Neural Networks

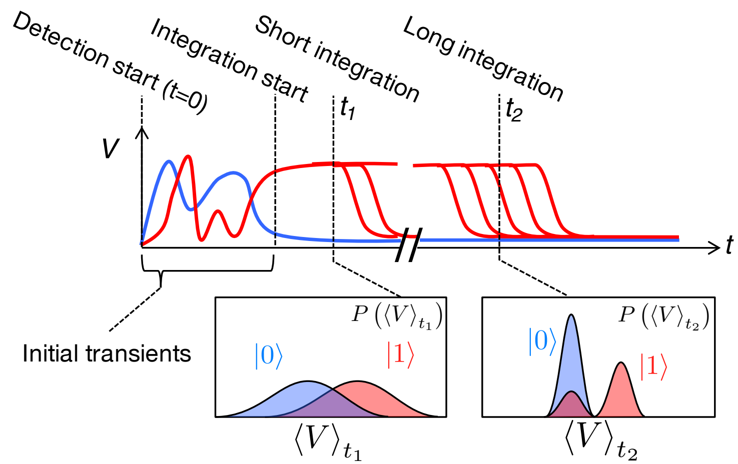
- **Nonlinear programming:**
  - Nonlinear optimization problems (robotics, predictive control, autonomous vehicles)
  - Partial differential equations
- **High-performance computing and Machine Learning:**
  - Vector-matrix multiplications
  - Deep learning inference
  - Ultrafast and online learning
- **Physics**
  - Qubit readout classification
  - High-energy particle collision experiments
- **Intelligent signal processing**
  - Optical fiber communications
  - Wideband RF signals
  - Spectral mining

- **Predictive Control**



Minimizing error between reference trajectory and measured output

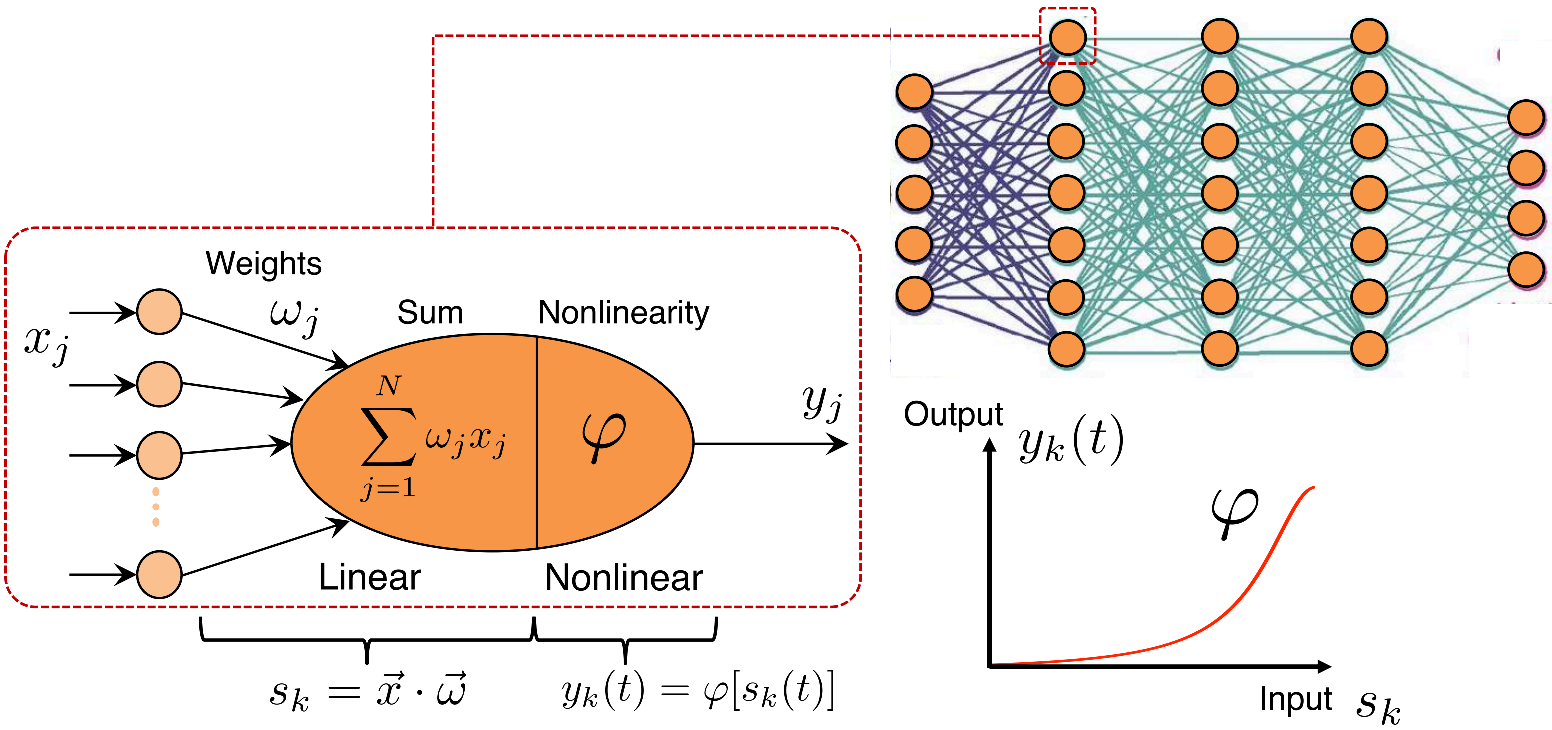
- **Qubit Readout Classification**



Tradeoff between low-SNR (short) and decay probability (long)
Turn-on transients are indicative of state

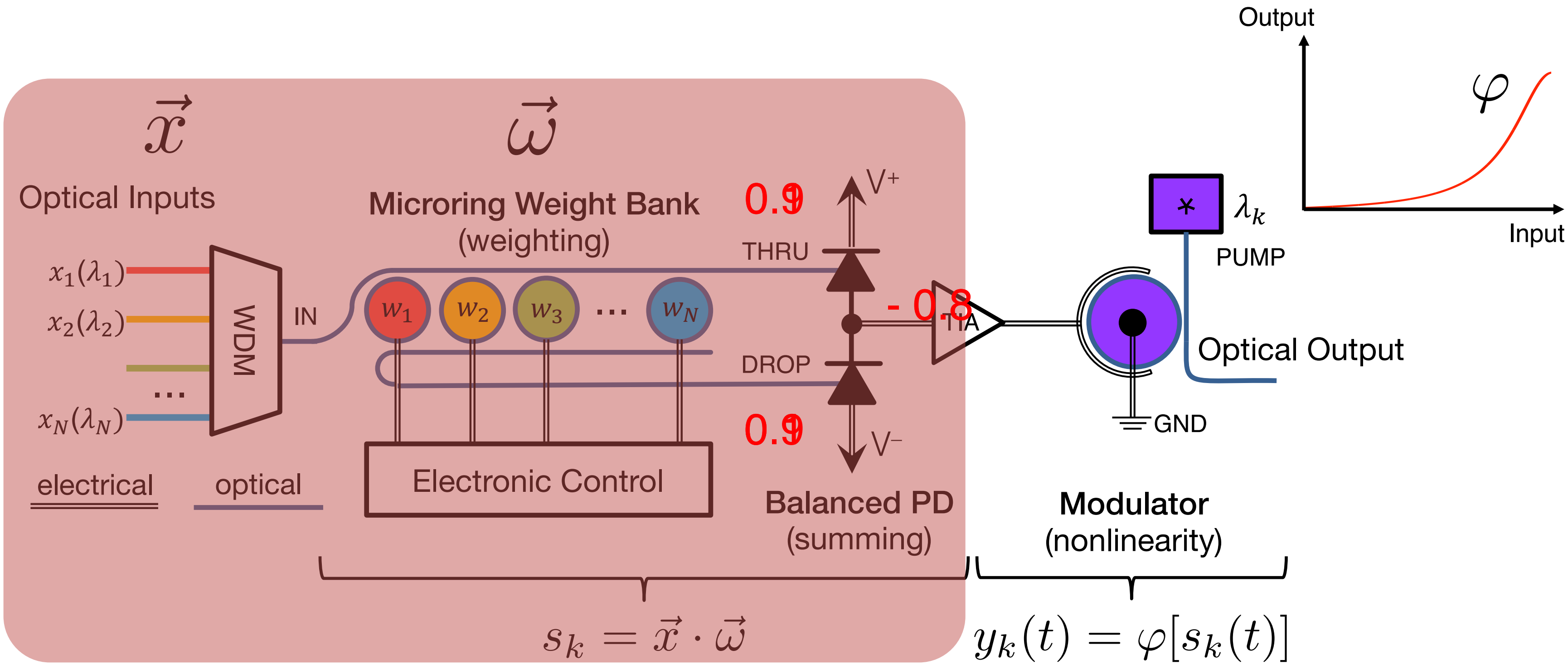
Shastri, Tait, Bhaskaran, Pernice, Prucnal et al. *Nature Photonics* 15 (2021)

Neuron Model and Weighted Sum



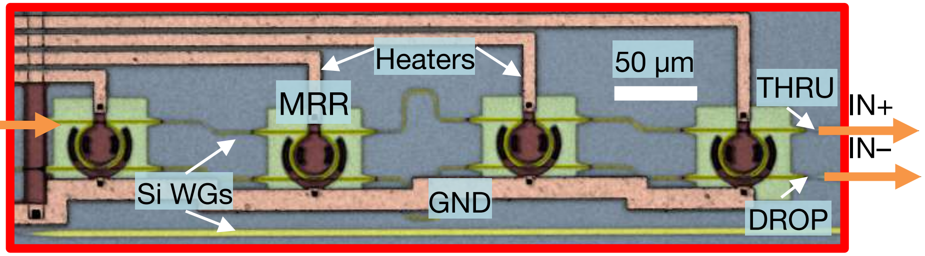
- **Neuron operation:** 1) receive multiple inputs; 2) weight each by coefficient; 3) sum them; 4) perform a nonlinear thresholding
- <u>Linear:</u> 1-3 is a weighted sum or **multiply & accumulate** (MAC)
- <u>Nonlinear:</u> state variable s is thresholded by a nonlinearity ϕ

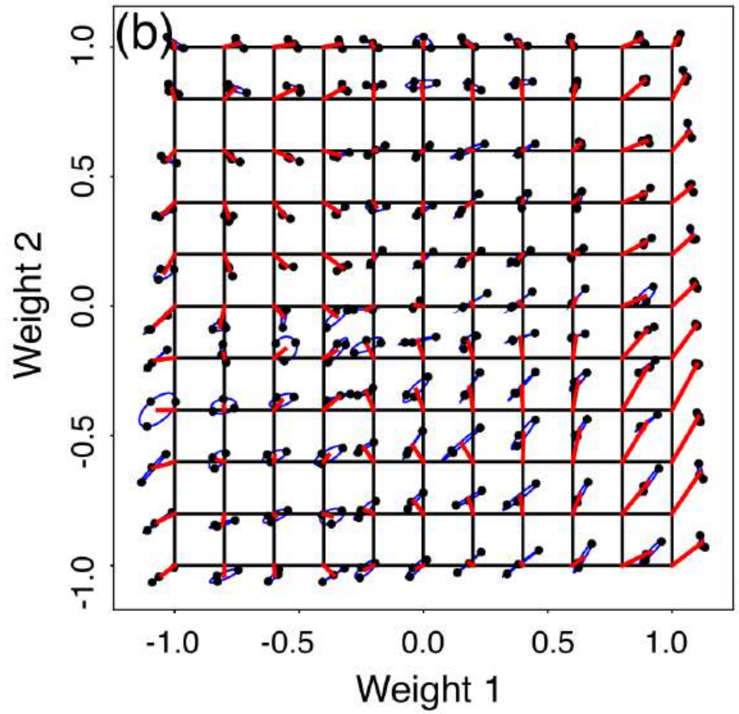
Neuron Model and Photonic Implementation



- **Neuron operation:** 1) receive multiple inputs; 2) weight each by coefficient; 3) sum them; 4) perform a nonlinear thresholding
- Linear: 1-3 is a weighted sum or **multiply & accumulate** (MAC)
- Nonlinear: state variable s is thresholded by a nonlinearity ϕ

Microring Weight Banks (Weighting)

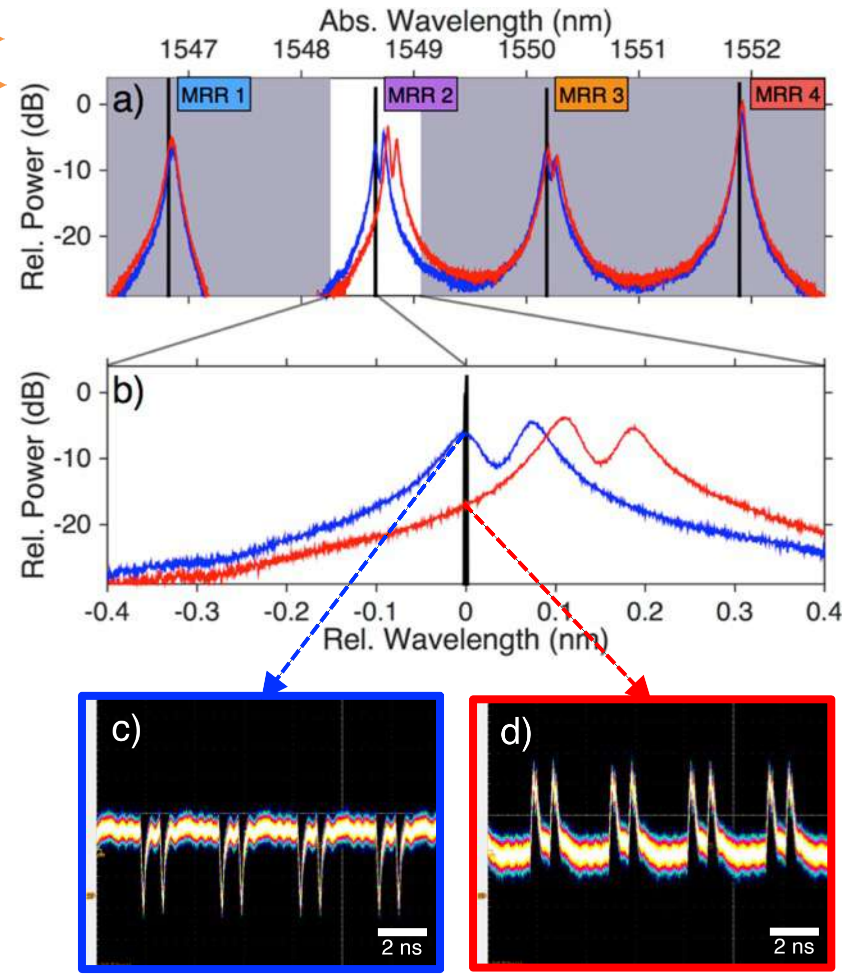




Red = mean error; Blue = std. dev.

Control precision: 7 bits + Sign bit

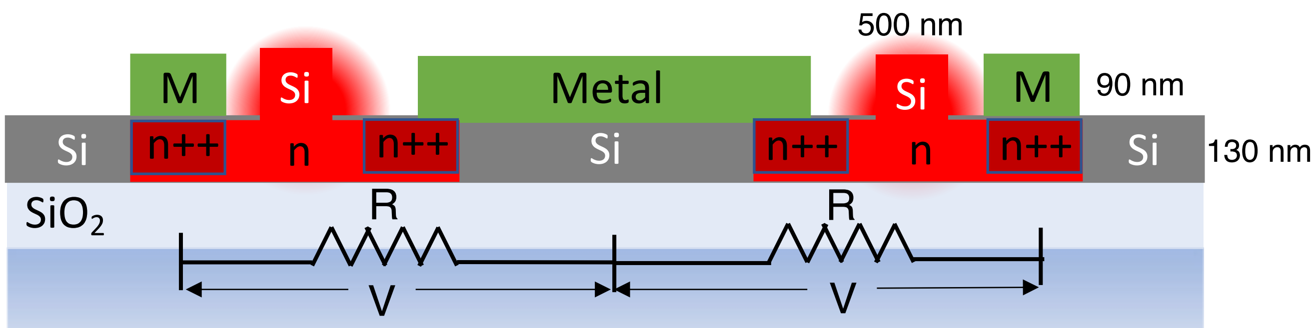
Measured Output of Drop Port



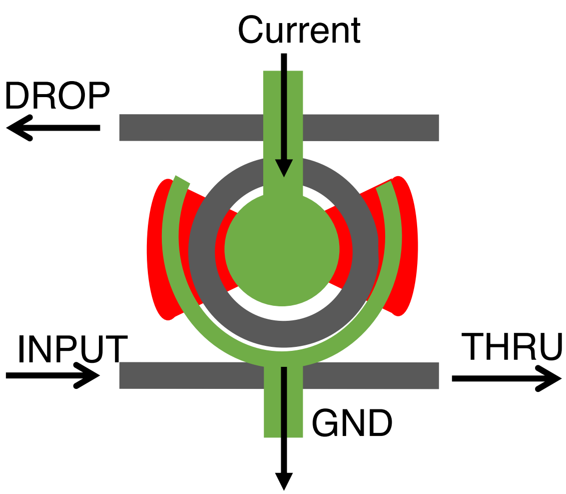
Balanced PD Output

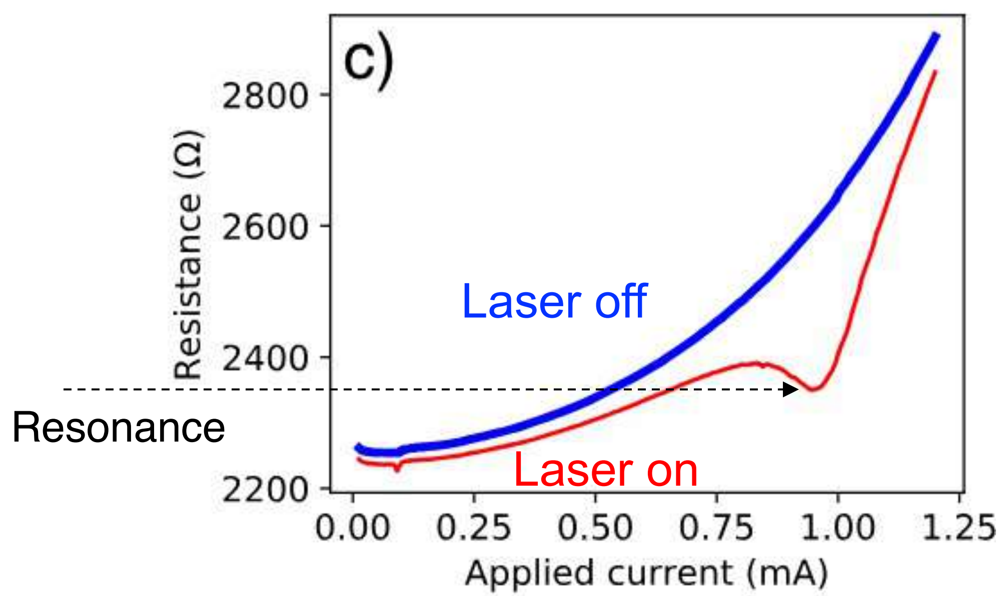
Huang, Shastri, Prucnal et al. *APL Photonics* **5** (2020)
Ma, Shastri, Prucnal et al. *Opt. Express* **27** (2019)

Silicon MRR Weight Bank with N-doped Heater



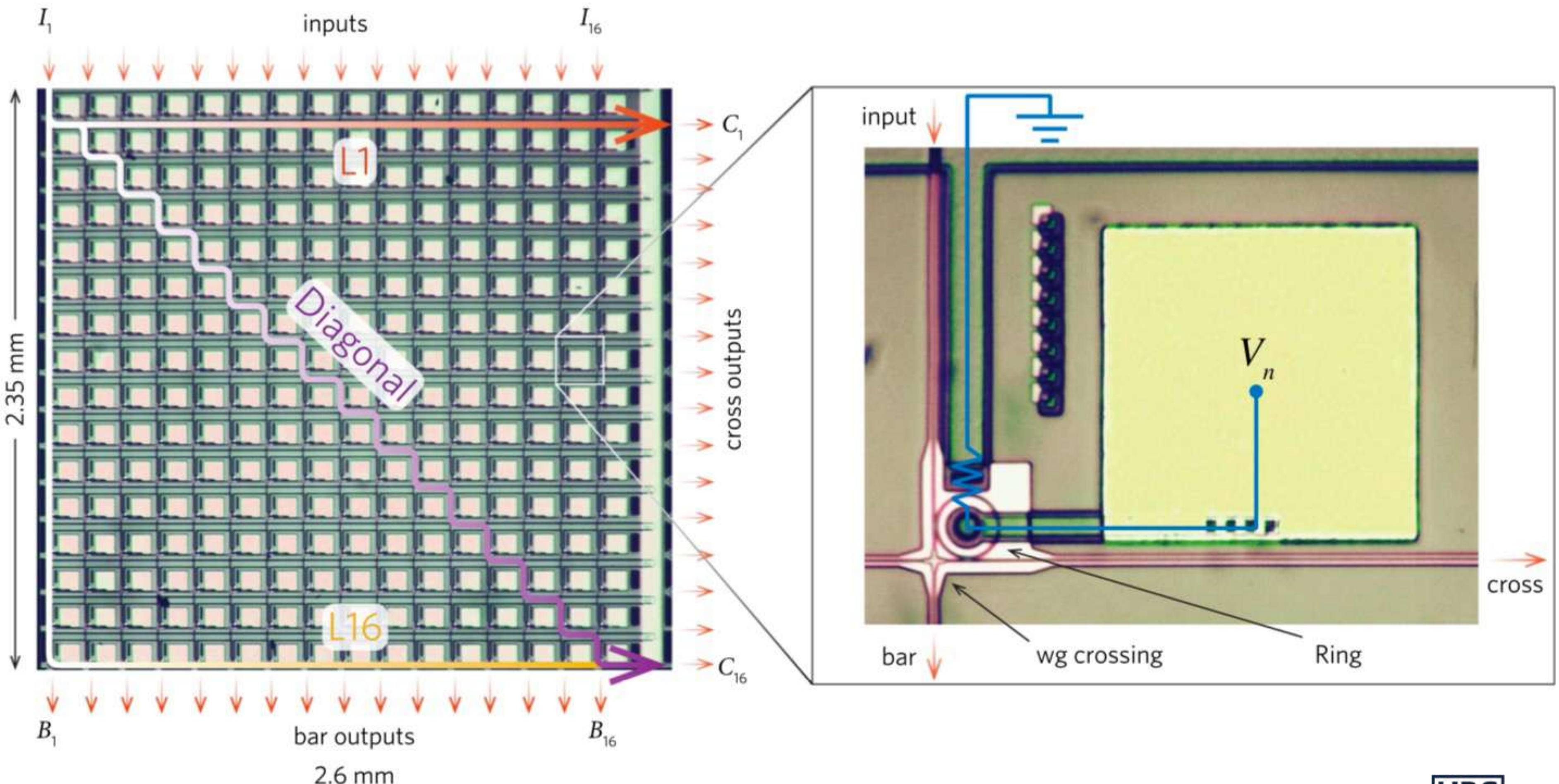
- Lower thermal crosstalk and more efficient heating
- Electrically sense the microring optical transmission
- Feedback for weight control
- **Control precision: 7 bits + sign bit**





Tait, Chrostowski, Shekhar, Prucnal et al. *Opt. Express* **26** (2018)

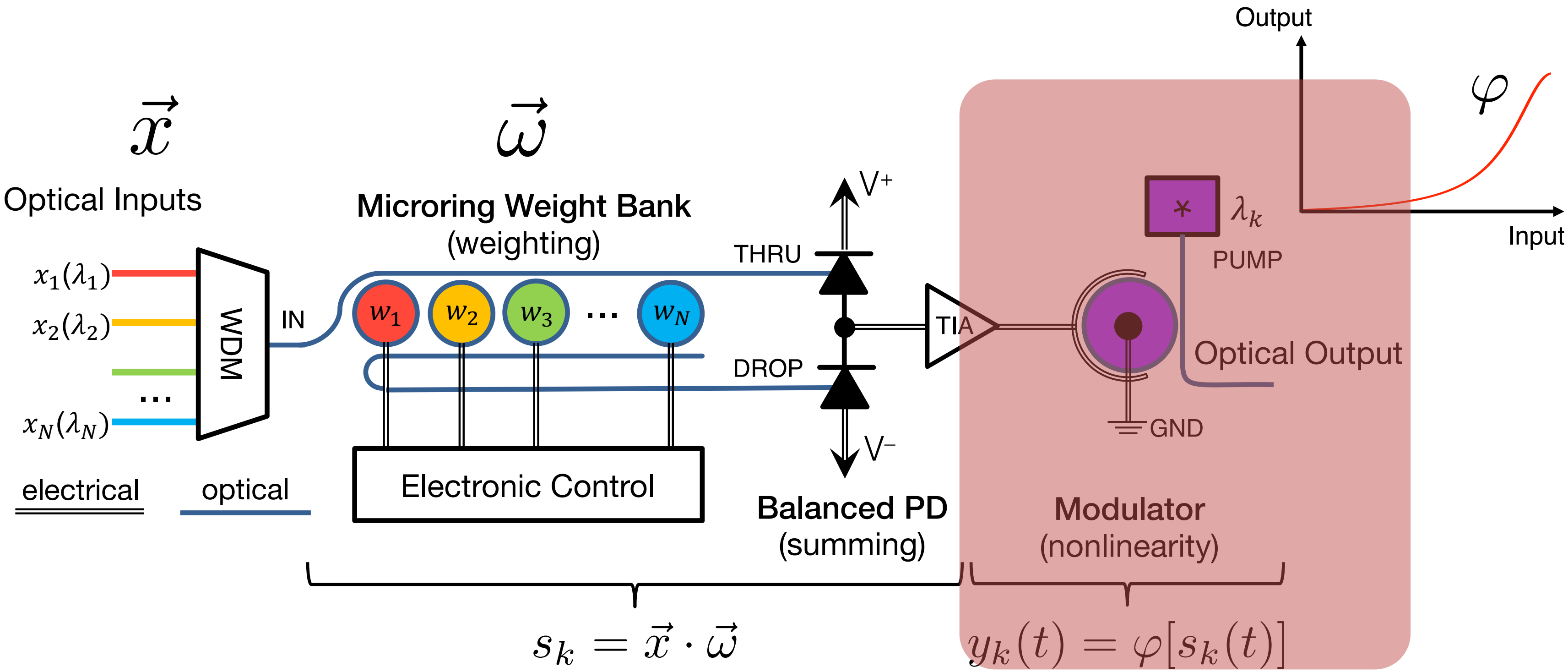
Large-Scale Control of MRRs



Collaboration: Lukas Chrostowski and Sudip Shekhar

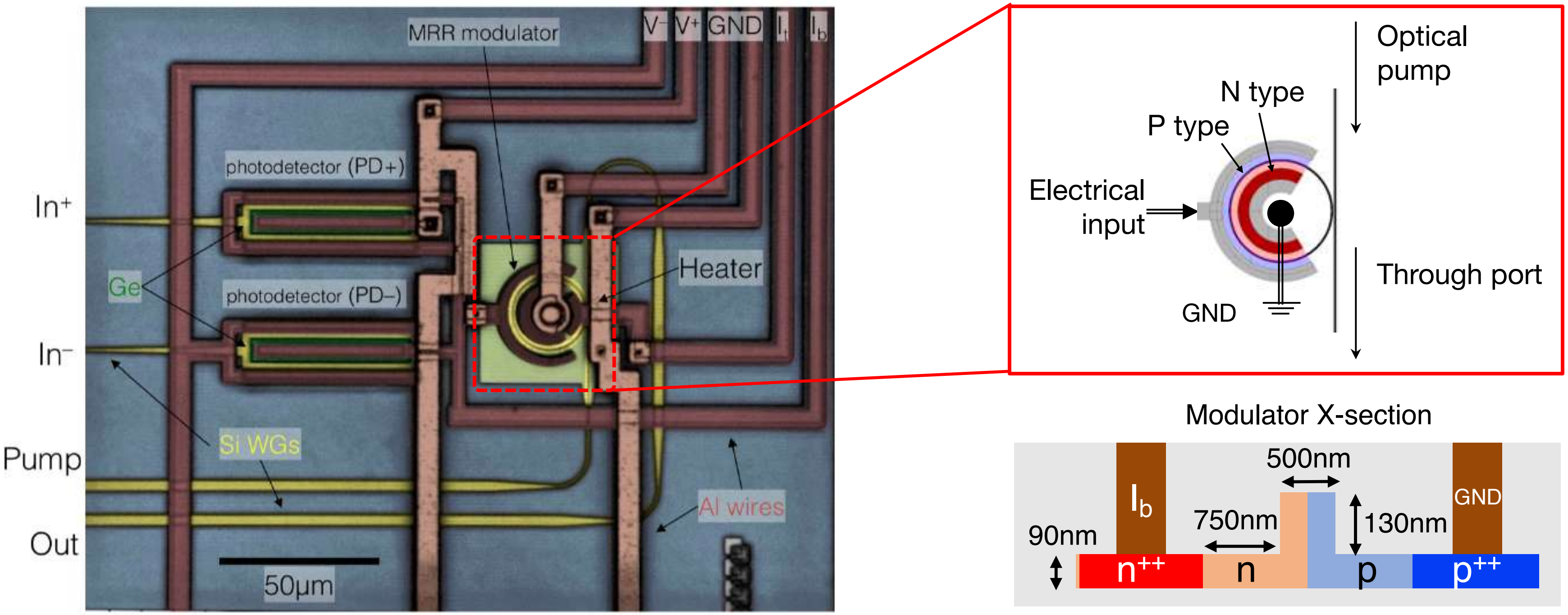
Jayatilleka, Chrostowski, Shekhar et al. *Optica* **6** (2019)

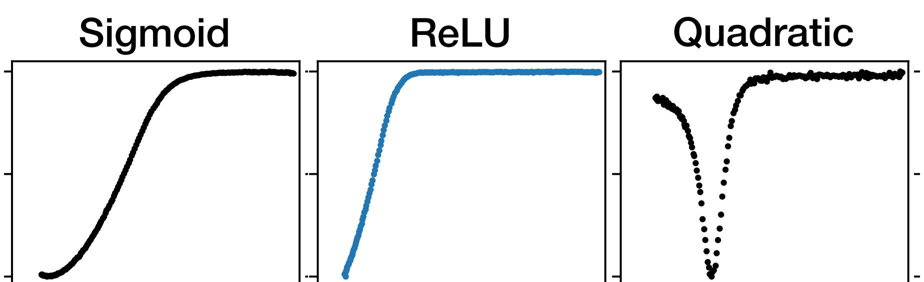
Neuron Model and Photonic Implementation



- **Neuron operation:** 1) receive multiple inputs; 2) weight each by coefficient; 3) sum them; 4) perform a nonlinear thresholding
- Linear: 1-3 is a weighted sum or **multiply & accumulate** (MAC)
- Nonlinear: state variable s is thresholded by a nonlinearity ϕ

Electro-Optic Nonlinearity: Reconfigurable Activation Functions





Recurrent Hopfield networks | *Feedforward (CNNs)* | *Support vector machines (SVMs)*

Modulate refractive index

$$\Delta n_{FC} = -8.8 \times 10^{-22}\Delta N - 8.5 \times 10^{-18}(\Delta P)^{0.8}$$

Modulate optical loss

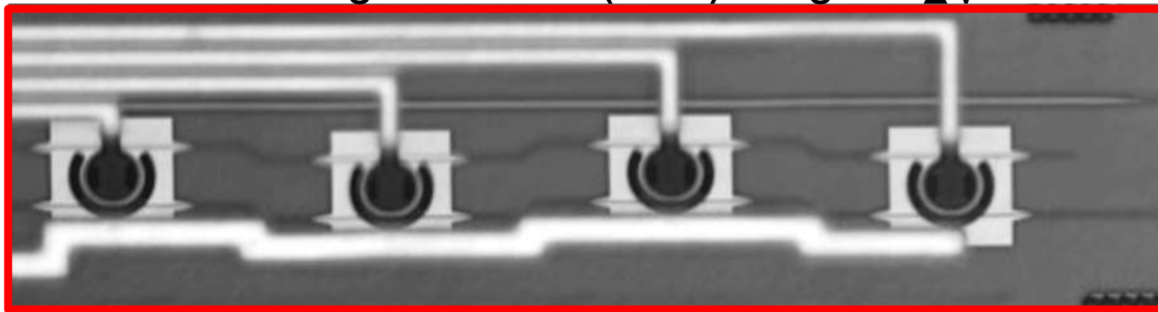
$$\alpha_{FCA} = (8.5 \times 10^{-18}\Delta N) + (6.0 \times 10^{-18}\Delta P)$$

- Bias current shifts the resonance and loss.
- Allows engineering the transfer function

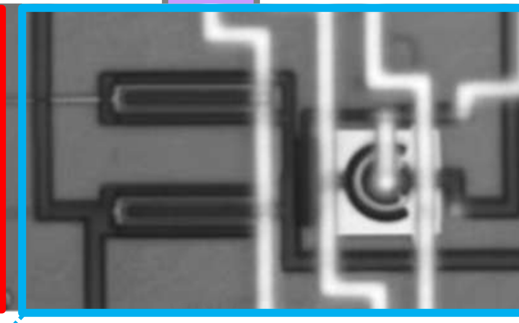
Tait, Shastri, Prucnal et al. *Phys. Rev. Appl.* **11** (2019)

Multiwavelength Silicon Photonic Neuron

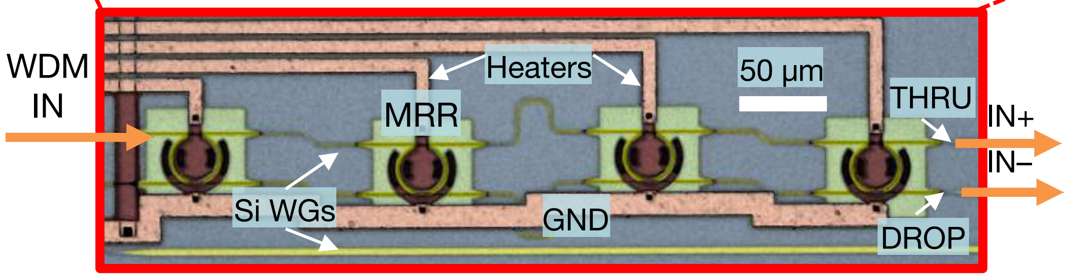
Microring Resonator (MRR) Weight Bank



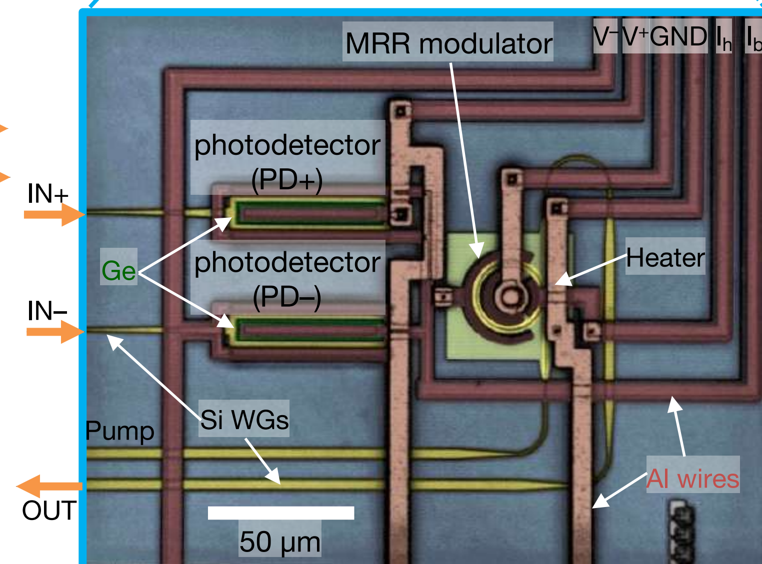
PD + Modulator



weighted addition (dot products): $s_k = \vec{x} \cdot \vec{\omega}$



nonlinearity: $y_k(t) = \varphi[s_k(t)]$



Features:

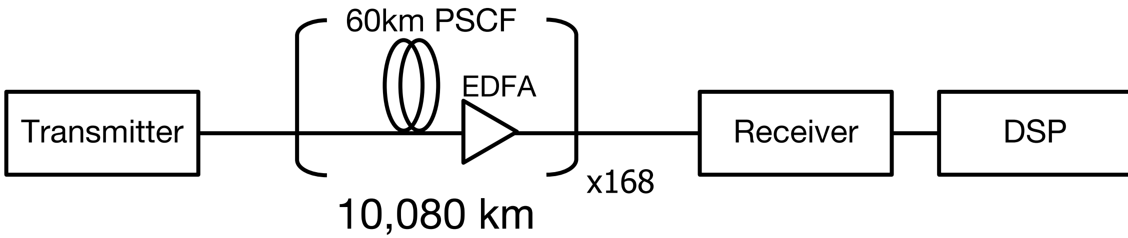
- MRR weight banks for WDM weighting
  - parallel processing & speed of light
- PD+/PD- do summing; allow for +ve/-ve weights
- MRR & balanced PD does vector multiplication
- Modulator for electro-optic nonlinearity
- Compatible with mainstream foundries

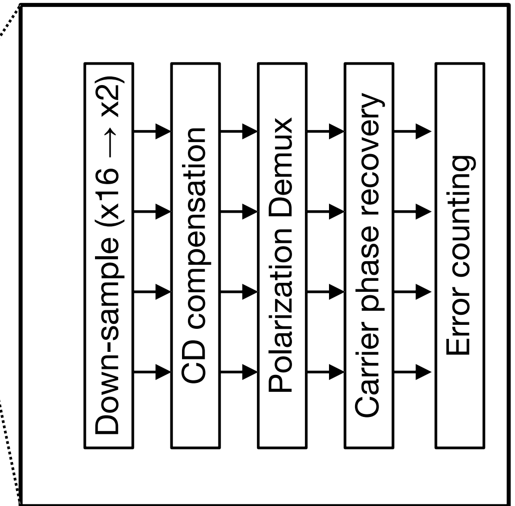
Metrics:

- Energy efficiency (today): 500fJ/MAC
- Efficiency (foreseeable tech.): 1.1 fJ/MAC
- Speed: 10 GHz operation

Tait, Shastri, Prucnal et al. *Phys. Rev. Appl.* **11** (2019)

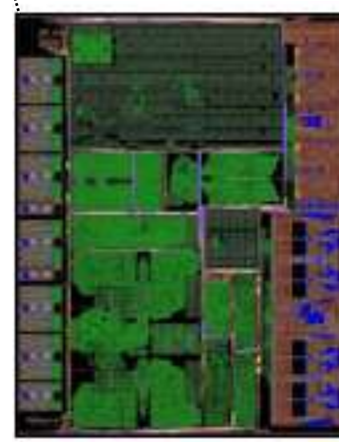
Need for AI in Fiber Optic Communications





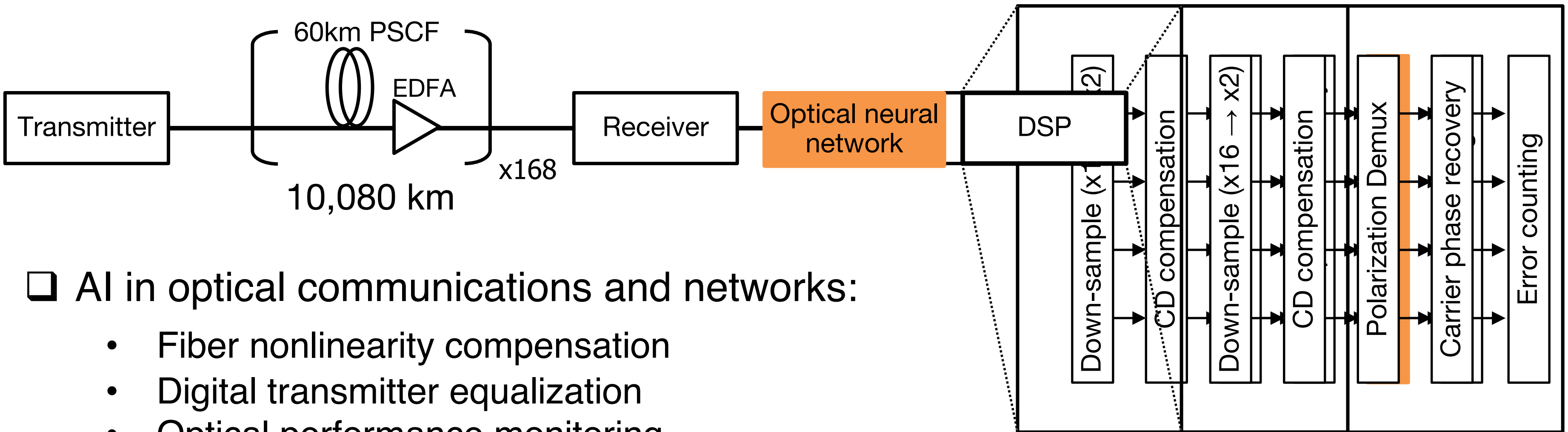
Constraints on DSP

- ❑ Increasing data rates (faster ADCs, power constraints)
- ❑ Increasing functionality:
  - Fiber nonlinearity compensation
  - Digital transmitter equalization
  - Optical performance monitoring
  - Traffic prediction
  - Failure management etc.
- ❑ Increases computational complexity: need for low latency, real-time processing

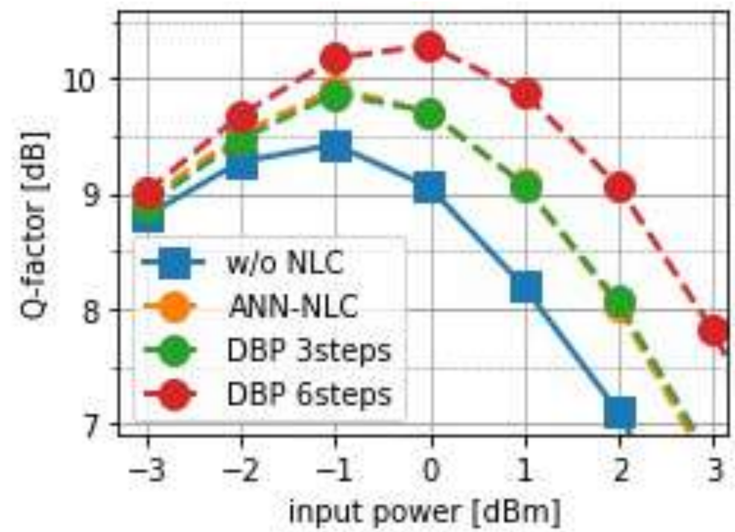


- 200G (14000km); 800G (200km)
- 7nm CMOS Process
- 800 Trillion operations/second
- 12mm x 16mm

Need for AI in Fiber Optic Communications



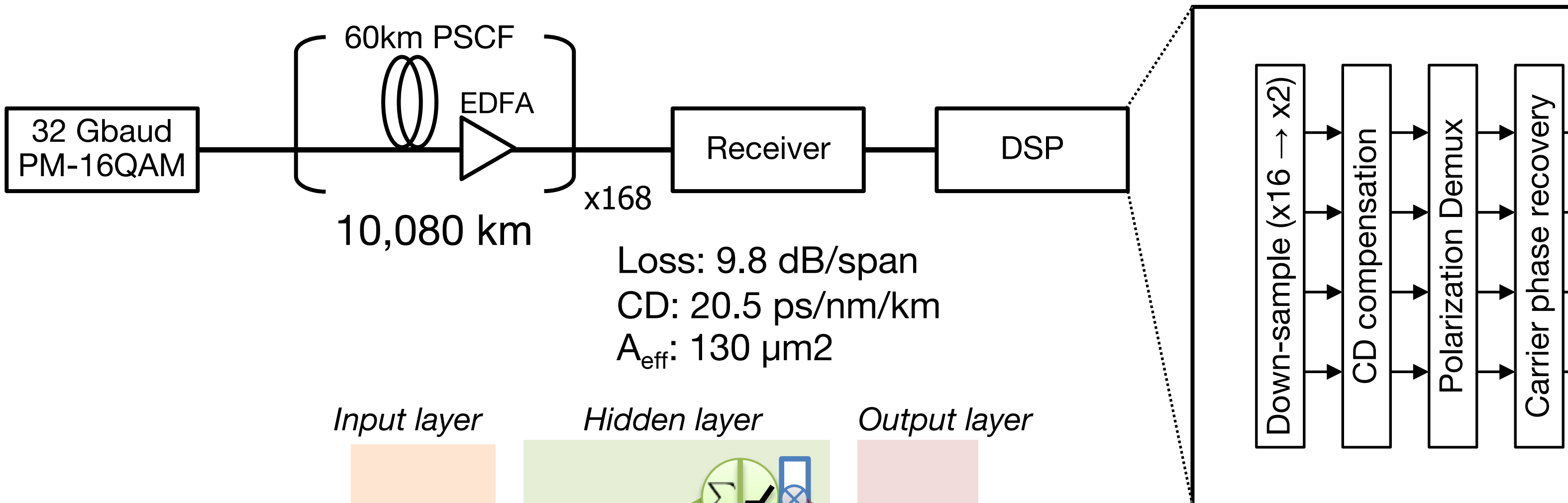
- AI in optical communications and networks:
  - Fiber nonlinearity compensation
  - Digital transmitter equalization
  - Optical performance monitoring
  - Traffic prediction
  - Failure management etc.
- AI vs. conventional approach:
  - ANN can learn nonlinear impairment without knowing transmission link parameters
  - Can outperform digital backpropagation (DBP) with low computational complexity
- Photonic neural networks:
  - Alleviate speed constraints
  - Process analog optical signals directly

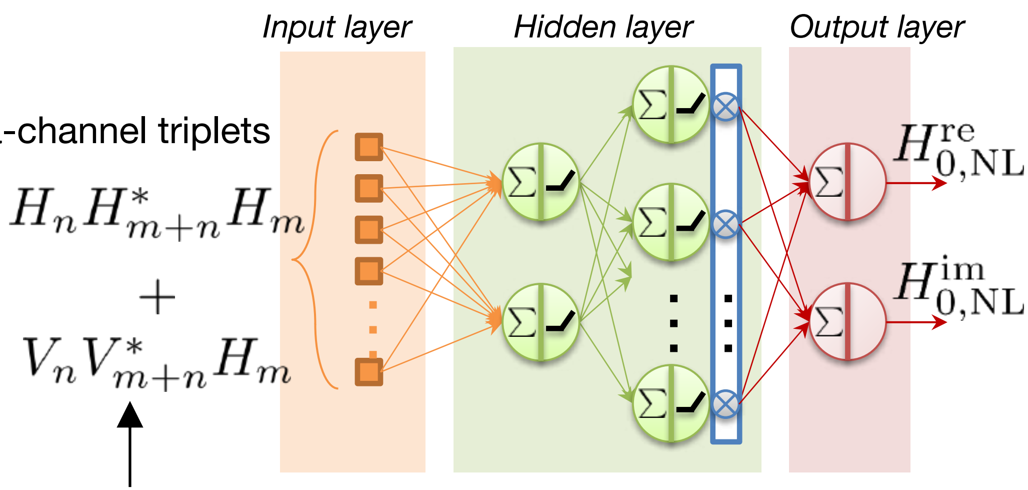


Zhang et al., *Nat. Commun.* **10**, 3033 (2019)

PRINCETON UNIVERSITY
NEC

Training ANN for Fiber Nonlinearity Compensation





- Intra-channel cross-phase (XPM) modulation and four-wave mixing (FWM)
- H (i.e. x)- and V (i.e. y) polarization
- VPItranmissionMaker simulation

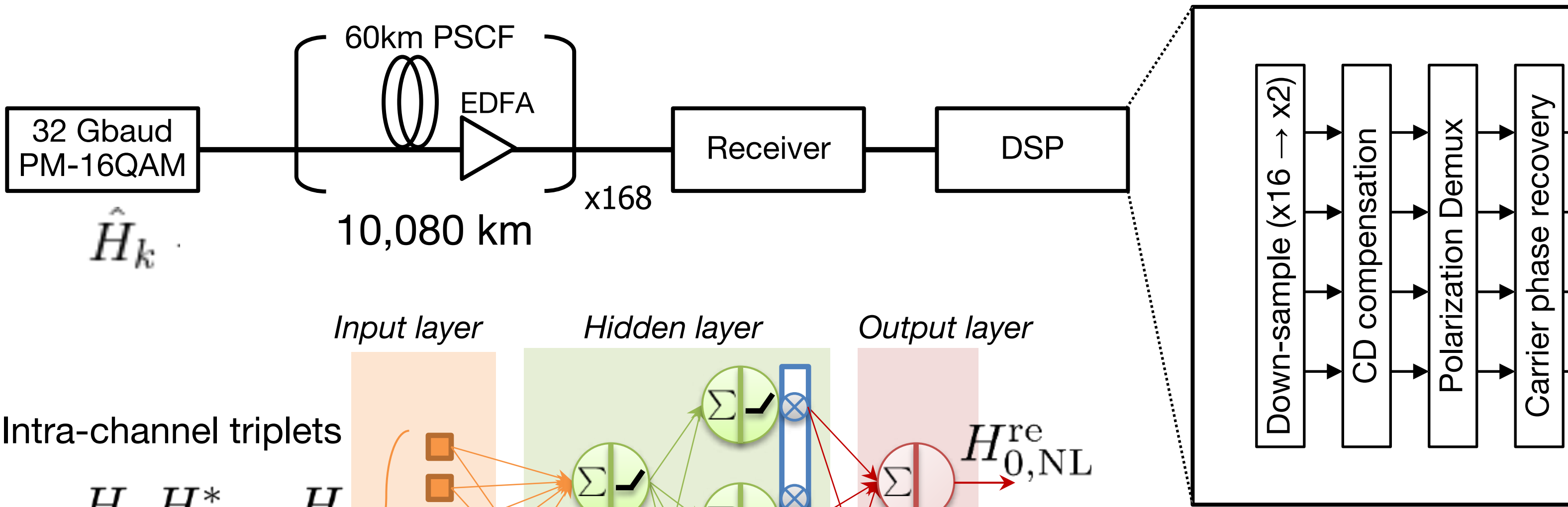
- 32 Gbaud PM 16 QAM
- 32106 training symbols
- 892 triplets
- 2 hidden layers
- 2 x 8 x 2 network
- Regression problem

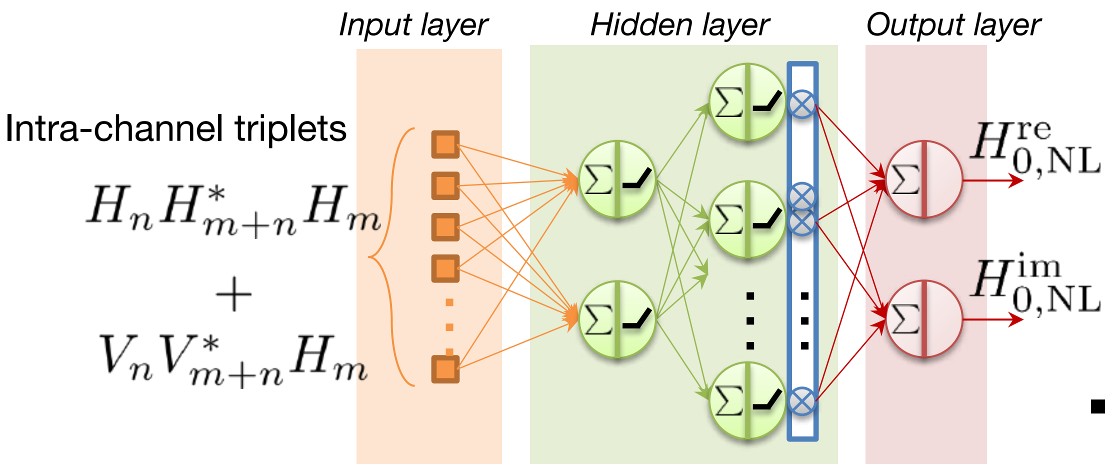
Huang*, Fujisawa*, Shastri, Wang, Prucnal et al. OFC 2020, Th4C.6

PRINCETON UNIVERSITY

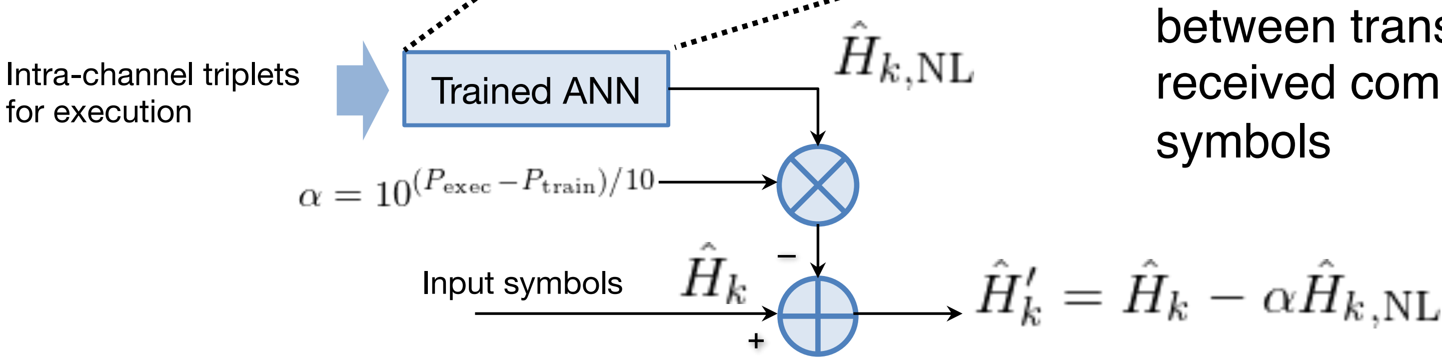
NEC

Training ANN for Fiber Nonlinearity Compensation





- Update weights and biases
- Minimize bit error rate between transmitted and received compensated symbols

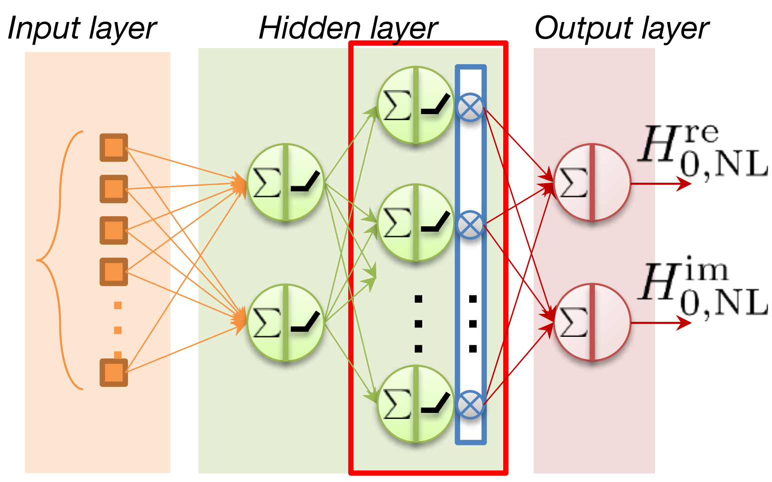


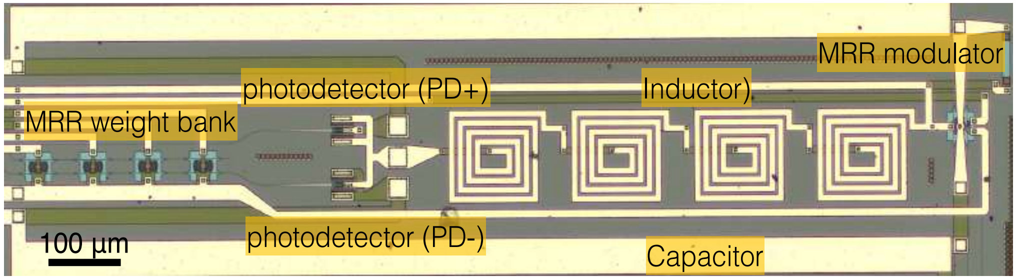
PRINCETON UNIVERSITY

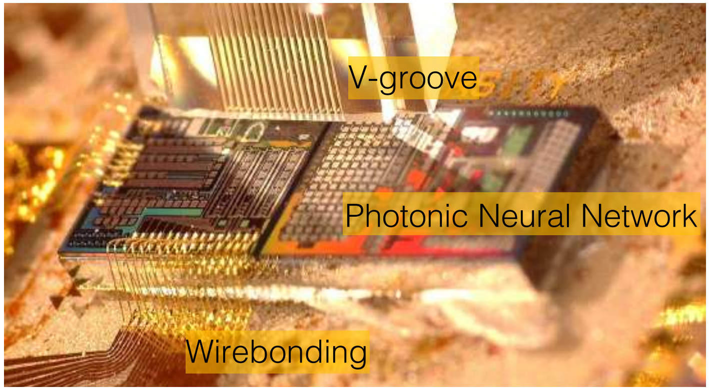
NEC

Huang*, Fujisawa*, Shastri, Wang, Prucnal et al. OFC 2020, Th4C.6

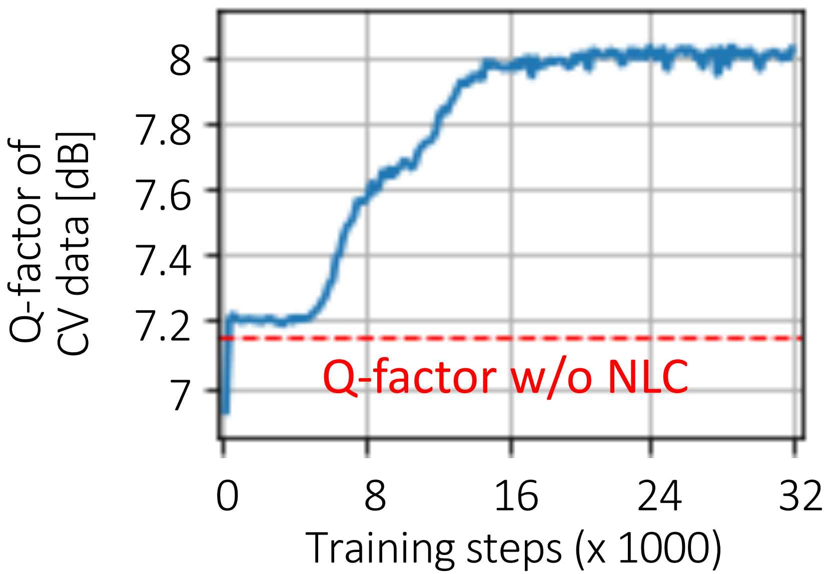
Neural Network Training using Lorentzian Function







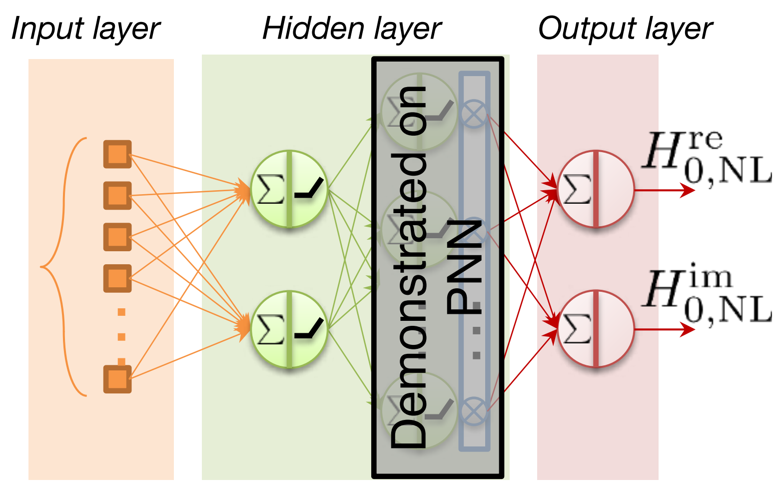
Neuron ID	weight 1	weight2	bias
1	0.0226	0.9997	0.0728
2	0.9898	0.1426	-0.0003
3	-0.1932	0.9812	0.2126
4	0.0516	0.9987	0.0645
5	0.9865	0.1637	-0.0019
6	-0.1736	0.9848	0.2057
7	0.9888	0.1492	0
8	0.9802	-0.1980	0.2338



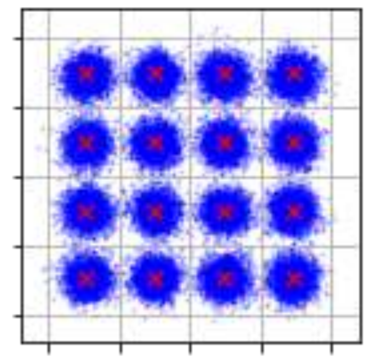
PRINCETON UNIVERSITY NEC

Nonlinear Compensation (NLC) Gain using PNN

- Evaluated the NLC gain of PNN-NLC
  - Q-factor of the transmission signals reconstructed from the eight photonic neuron outputs of the whole frame

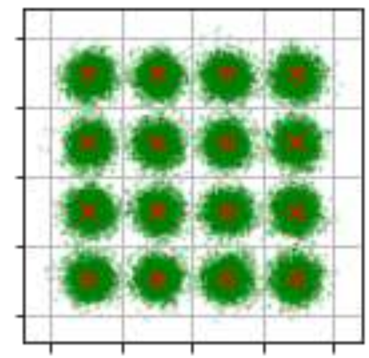


PM-16QAM (X-pol.)



ANN-NLC by <u>numerical</u> simulations

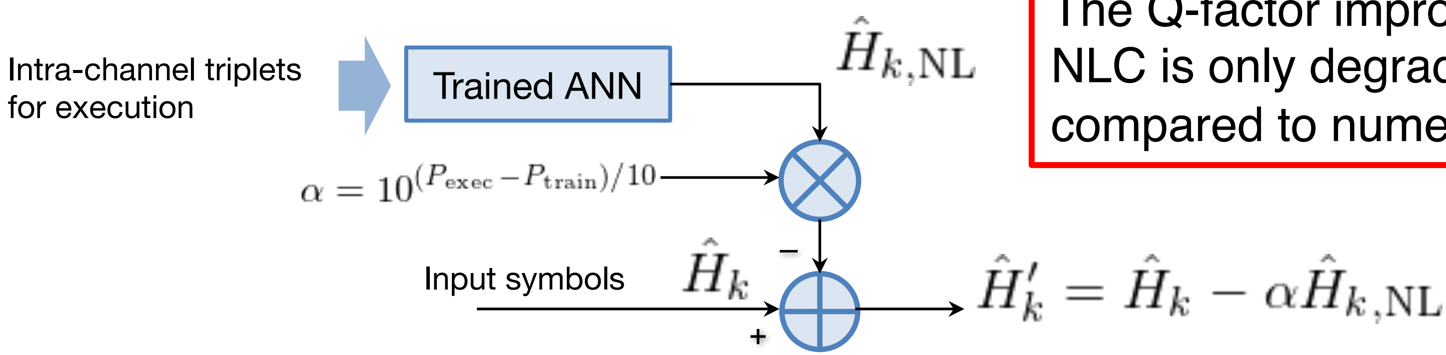
NLC gain 0.57 dB



PNN-NLC by <u>experimental</u> demonstration

NLC gain 0.51 dB

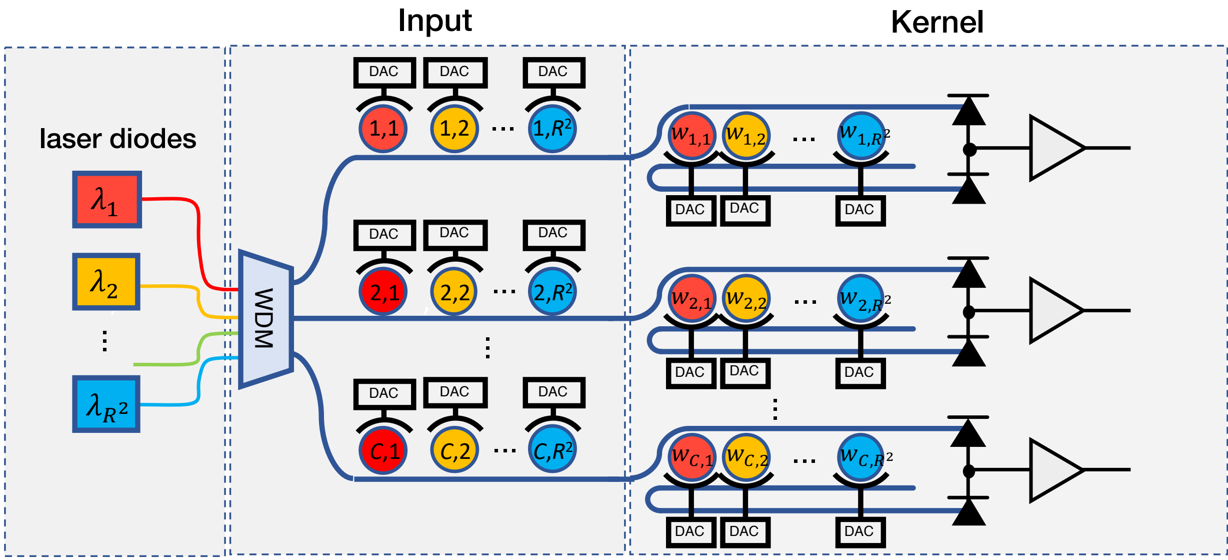
The Q-factor improvement of PNN-NLC is only degraded **by 0.06 dB** compared to numerical simulation

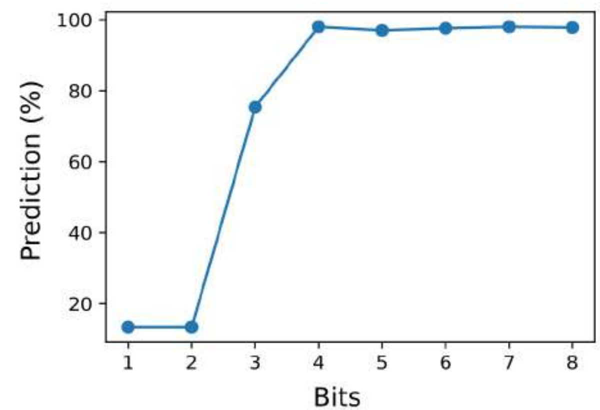


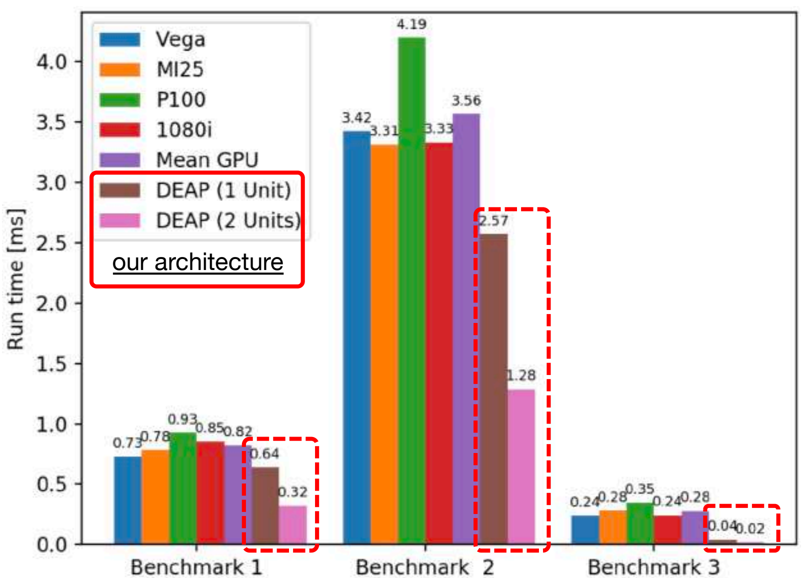
PRINCETON UNIVERSITY

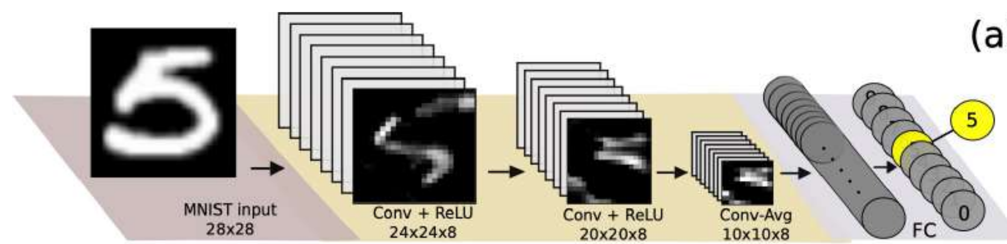
NEC

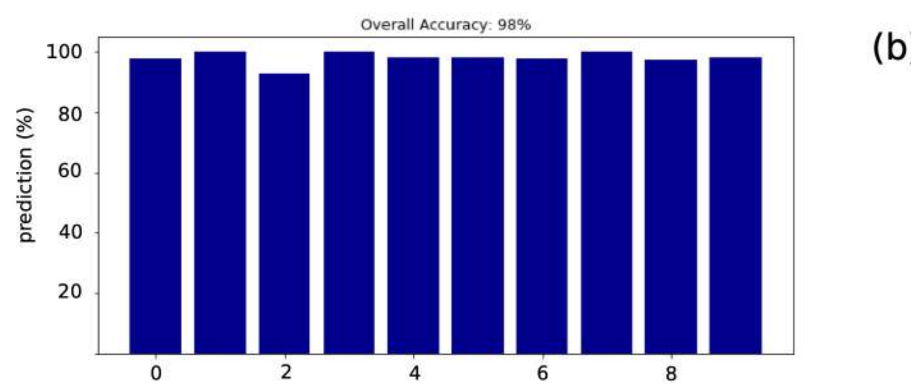
Photonic Convolution Accelerator for Deep Learning







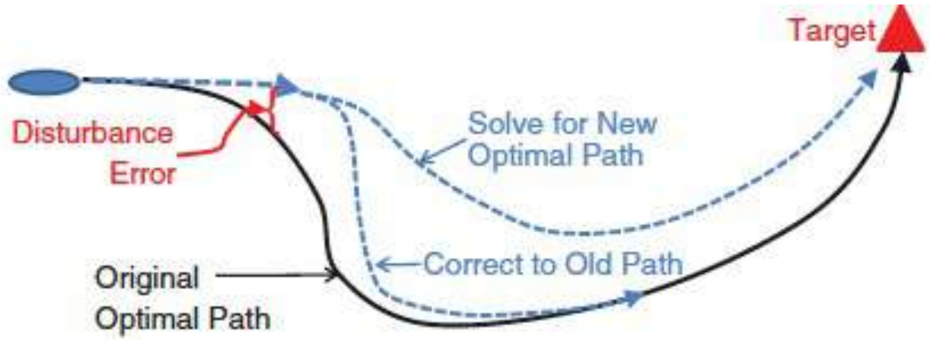




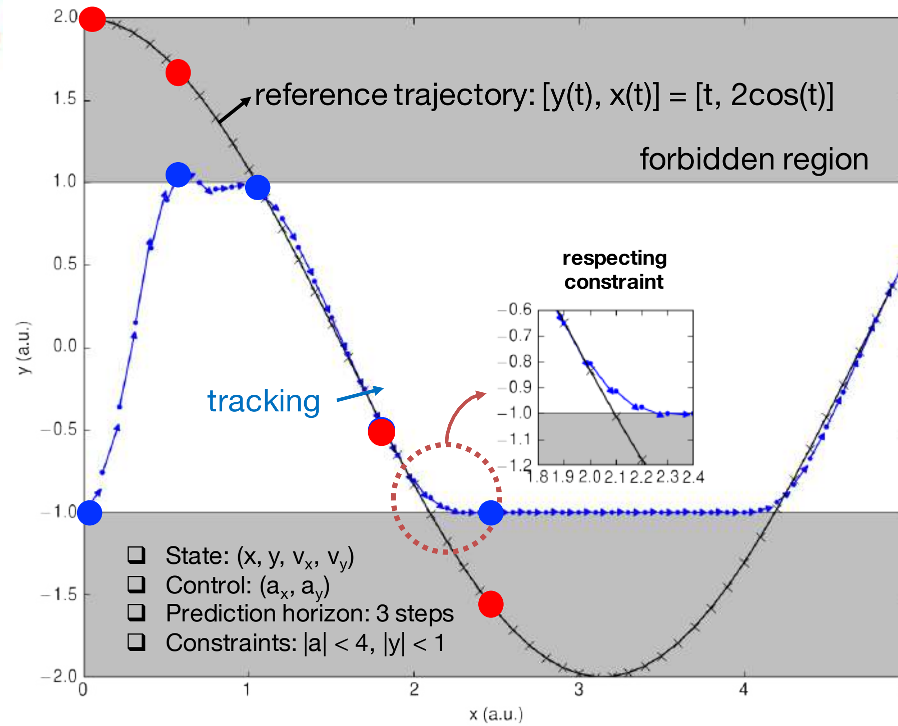
DEAP: https://github.com/Shastri-Lab/DEAP

Bangari, Prucnal, Shastri et al. *IEEE JSTQE* **26** (2020)

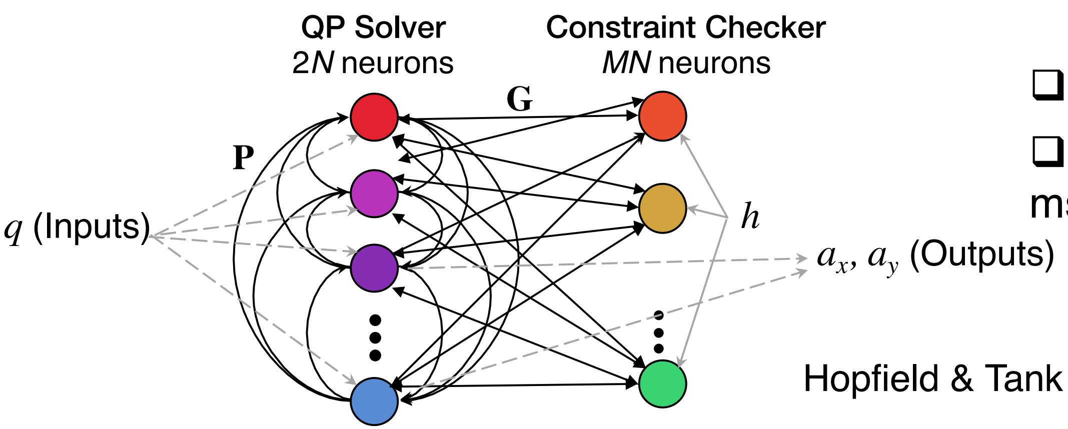
Nonlinear Programming: Predictive Control



Minimizing error between reference trajectory and measured output



$$\min_{x} J(x) = \frac{1}{2}\vec{x}^T \mathbf{P}\vec{x} + \vec{q}^T \vec{x} \text{ s.t. } \mathbf{G}x \leq \vec{h}$$

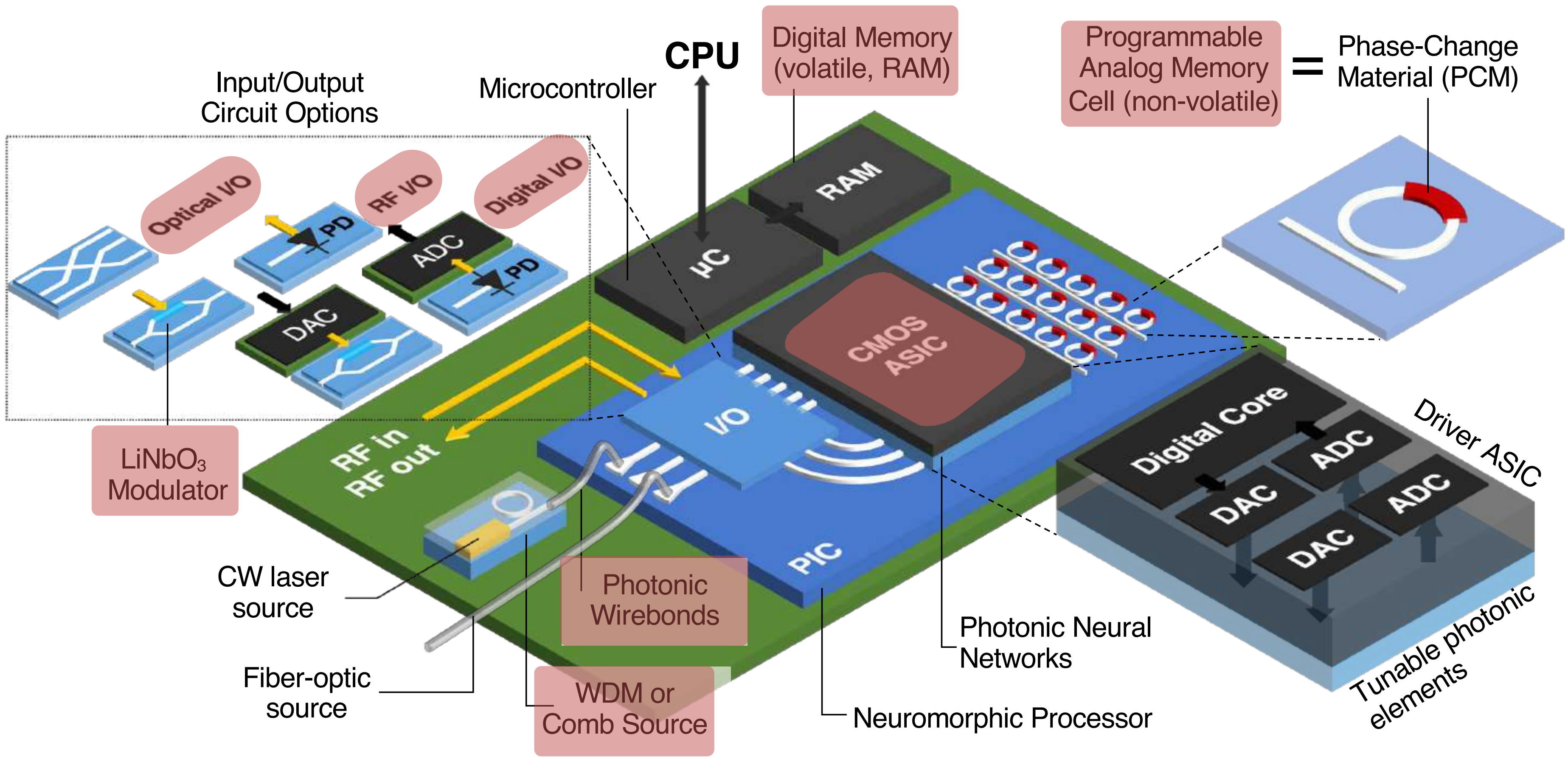


- convergence time in the order of 10 ns
- electronic controllers solving MPC (~10 ms)

Hopfield & Tank *Science* **233** (1986)

Ferreira de Lima, Shastri, Prucnal et al. *J. Light. Technol.* ***37*** (2019)

Roadmap: Emerging Ideas and Challenges



Shastri, Tait et al. *Nature Photonics* **15** (2021)

REVIEW ARTICLE | FOCUS
https://doi.org/10.1038/s41566-020-00754-y
nature photonics

Photonics for artificial intelligence and neuromorphic computing

Bhavin J. Shastri[1,2,7] ✉, Alexander N. Tait[2,3,7] ✉, T. Ferreira de Lima[2], Wolfram H. P. Pernice[4], Harish Bhaskaran[5], C. D. Wright[6] and Paul R. Prucnal[2]

Research in photonic computing has flourished due to the proliferation of optoelectronic components on photonic integration platforms. Photonic integrated circuits have enabled ultrafast artificial neural networks, providing a framework for a new class of information processing machines. Algorithms running on such hardware have the potential to address the growing demand for machine learning and artificial intelligence in areas such as medical diagnosis, telecommunications, and high-performance and scientific computing. In parallel, the development of neuromorphic electronics has highlighted challenges in that domain, particularly related to processor latency. Neuromorphic photonics offers sub-nanosecond latencies, providing a complementary opportunity to extend the domain of artificial intelligence. Here, we review recent advances in integrated photonic neuromorphic systems, discuss current and future challenges, and outline the advances in science and technology needed to meet those challenges.

Shastri, Tait et al. *Nature Photonics* **15** (2021)

Several Startups with Real Products

LUMINOUS

LIGHTELLIGENCE

LIGHTMATTER

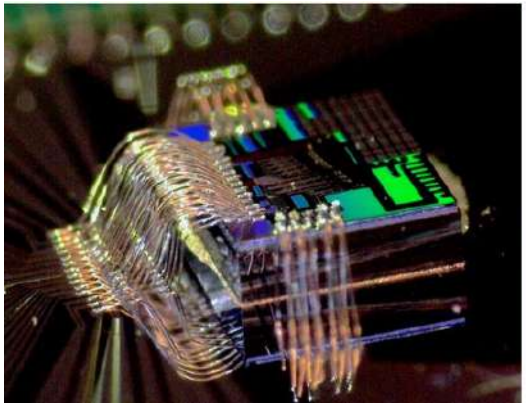
OPTELLIGENCE

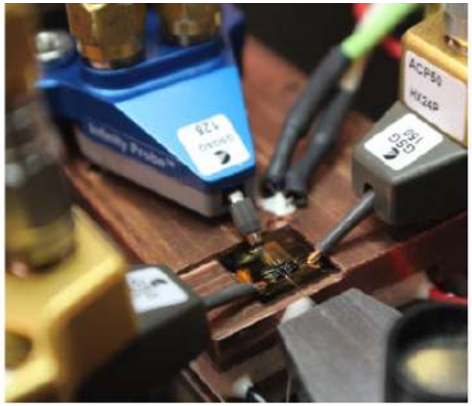
LightOn

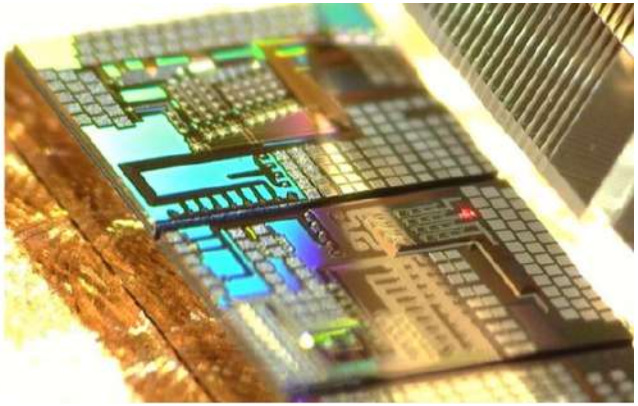
OPTIUS

Conclusion

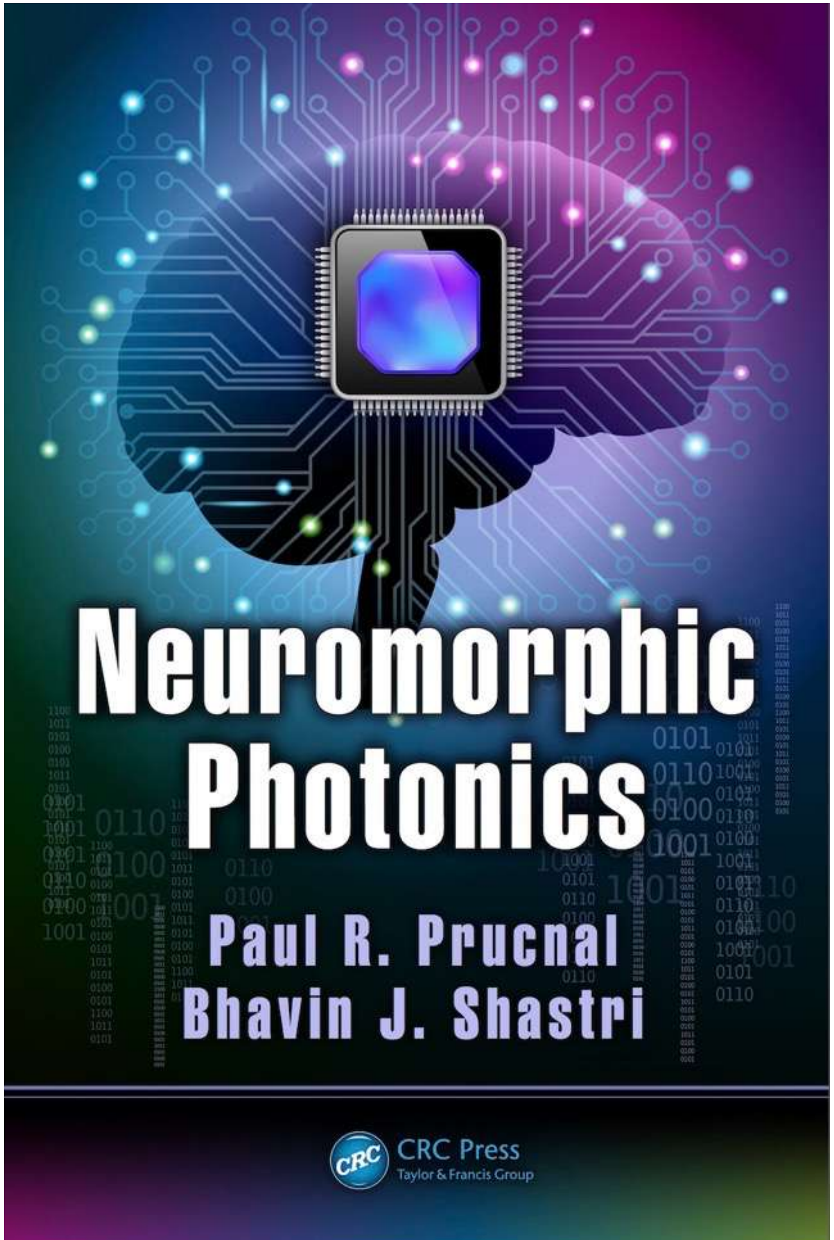
- Photonic primitives and integrated network:
  - Carrier & photon dynamics allow very fast processing (ps versus ms)
  - Overcome the bandwidth limits of electronic fan-in & interconnection
- Wavelength-based networking provides massive interconnection
- Compatible with commercial integration technology
  - Implementation on an SOI platform may allow for a scalable photonic computing system
- Photonic networks good for performing same task over and over quickly
  - Machine learning: convolutions for deep learning architectures
  - Neuromorphic computing: solving nonlinear optimization problems







Book



- Taylor & Francis (CRC Press)
- 412 Pages
- ISBN 9781498725224
- Contributions:

  Tait, Ferreira de Lima, Nahmias