JANUARY 2025

TEXAS
The University of Texas at Austin

A Compact Optical Neuron Based on Multi-Operand Microring Resonators

Shupeng Ning,[a] Chenghao Feng,[a] Jiaqi Gu,[a,b] Hanqing Zhu,[a]

Rongxing Tang,[a] David Z. Pan,[a] and Ray T. Chen[a]

[a] Department of Electrical and Computer Engineering, The University of Texas at Austin;

[b] School of Electrical, Computer and Energy Engineering, Arizona State University

This work is supported by U.S. Air Force, MURI

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

The University of Texas at Austin

SPIE. PHOTONICS WEST

Outline

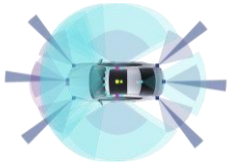
1. Background of Photonic AI
2. Hardware-efficient design for Optical Neural Networks
3. MRR-based Multi-Operand Optical Neuron (M^2OON)
4. Experimental Results of M^2OON with Tunable Nonlinear Active Functions
5. Summary

Photonic AI keeps growing

Why Photonic AI ?

- High computation speed and low power consumption
- High bandwidth in the **analog** domain
- Unique multiplexing techniques, *e.g.*, **WDM**
- Integrated photonics bring new opportunities…

AI influences various aspects of our lives

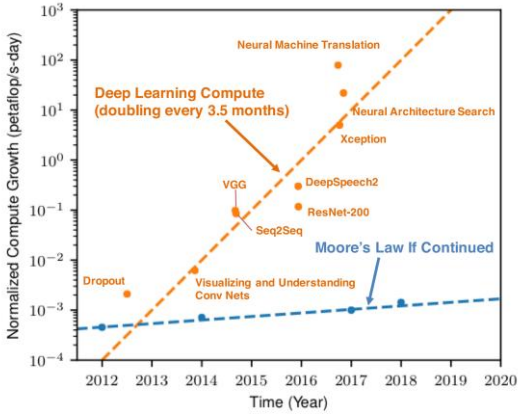


Autonomous driving

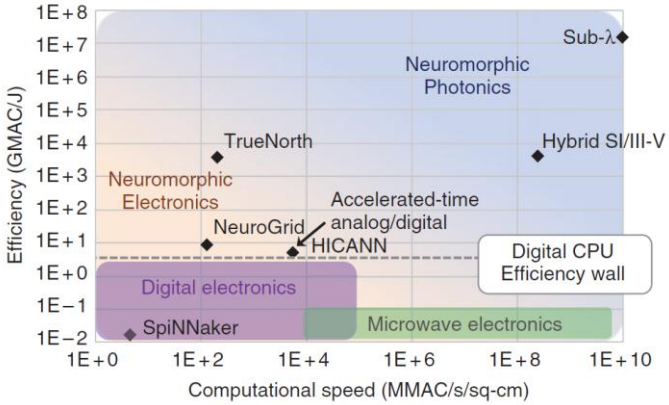
Data centers

ChatGPT

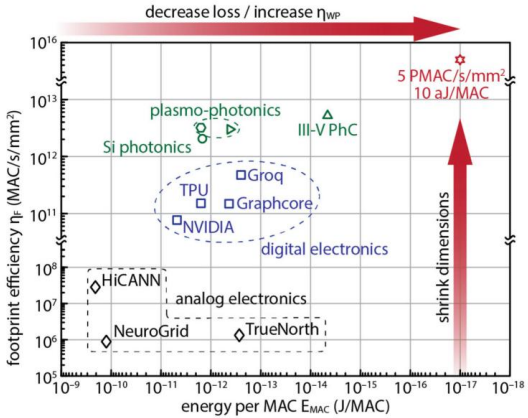
General Language Model



[Lima+, 2020]



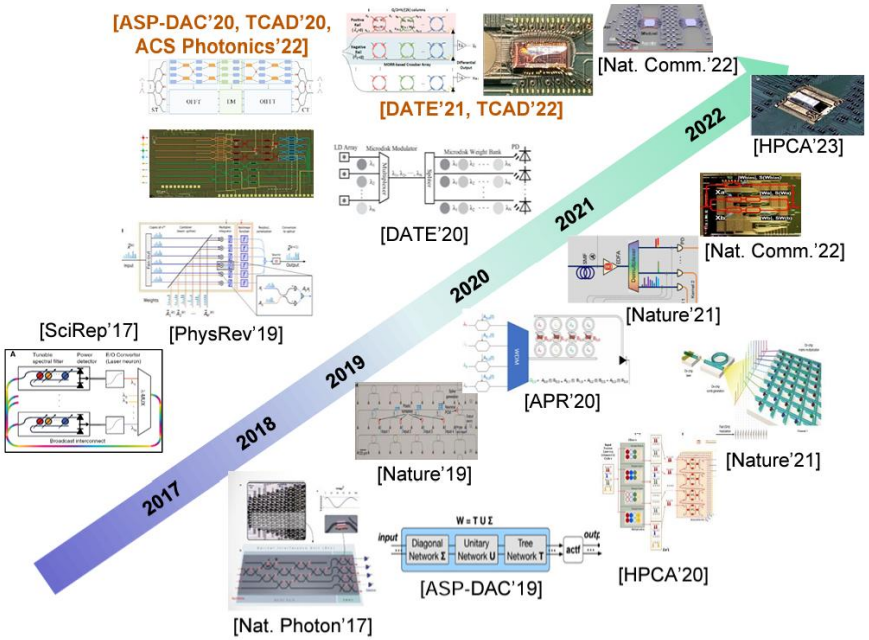
[Lima+, 2017]



[Totovick+, 2017]

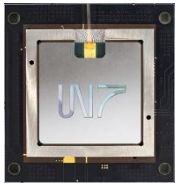
Photonic AI keeps growing

Photonic Neural Network Trends in Academia



Foundry / EPDA Support in Industry

Photonic Computing Chip Designs



LIGHTMATTER
LIGHTELLIGENCE
SALIENCE LABS
LUMINOUS
OPTIUS light powered computing
Lighton

Design Automation / Simulation Tools

Ansys LUMERICAL
SYNOPSYS OPTICAL & PHOTONIC SOLUTIONS
cadence
COMSOL

PDKs / Foundry

amf ADVANCED MICRO FOUNDRY
AIM PHOTONICS
SiEPIC FAB OpenEBL
tsmc
GlobalFoundries

Challenges and Concerns:

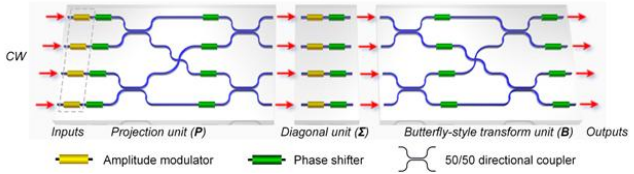
- Large footprints of optical devices in PICs → **Scalability**
- Inefficient electrical-optical (E-O) interfaces → **Efficiency**
- Training for hardware-based ONNs…

The absence of on-chip nonlinearities, which are critical for NNs, necessitates additional E-O/O-E conversion!

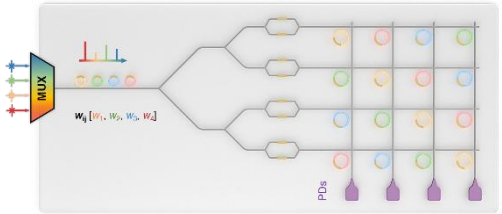
How to improve?

Toward Hardware-Efficient Optical Neural Networks (ONNs)

Circuit-level optimization

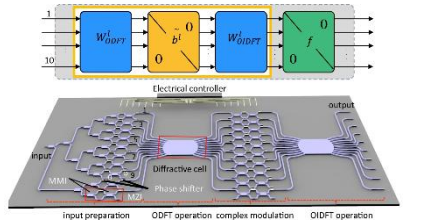


Butterfly-style photonic mesh [Feng+, 2022]

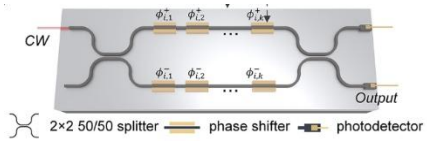


Structured compression approach [Ning+, 2024]

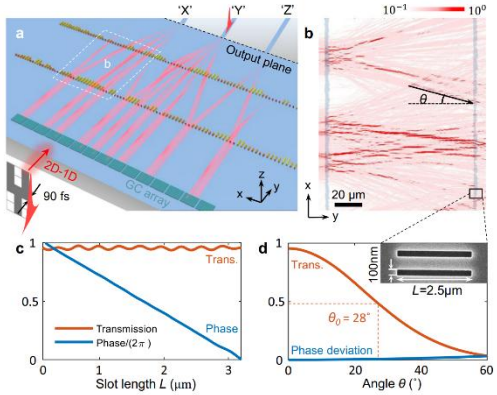
Deployment of compact device-level tensor cores



Star-coupler-based ONN [Zhu+,2022]



Multi-operand MZI [Feng+,2022]



Metasurface-based ONN [Wang+,2024]

Multi-Operand Modulators — *Squeeze operation on single device*

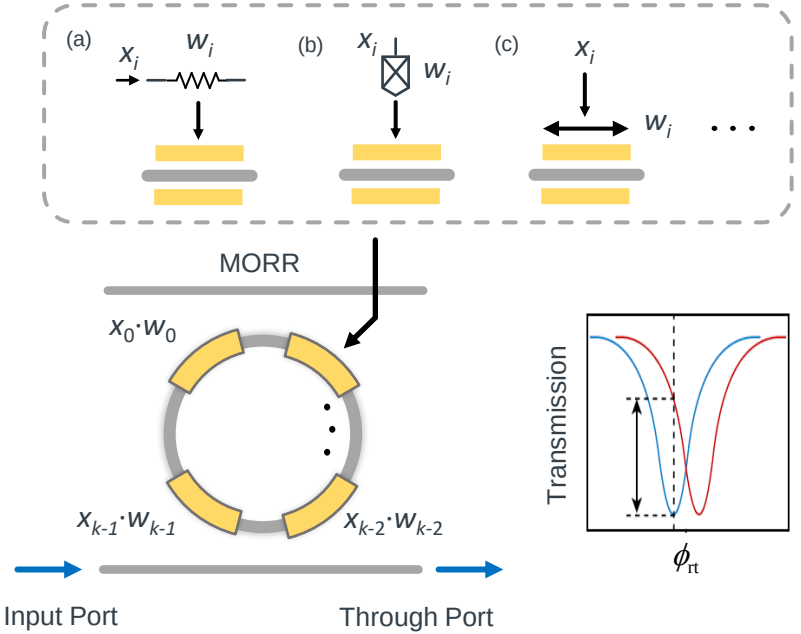
Multi-Operand Micro-ring Resonator (MORR)

- MORR has ***k*** active actuators controlled simultaneously by independent electrical signals.
- Squeezes a length-*k* tensor operation within a single device.

  k ×area/power/wavelength saving than MRR arrays
- Multiplication could achieve by (a) programmable resistors, (b) tunable amplifiers/attenuators, (c) adjusting modulation length, *etc*.

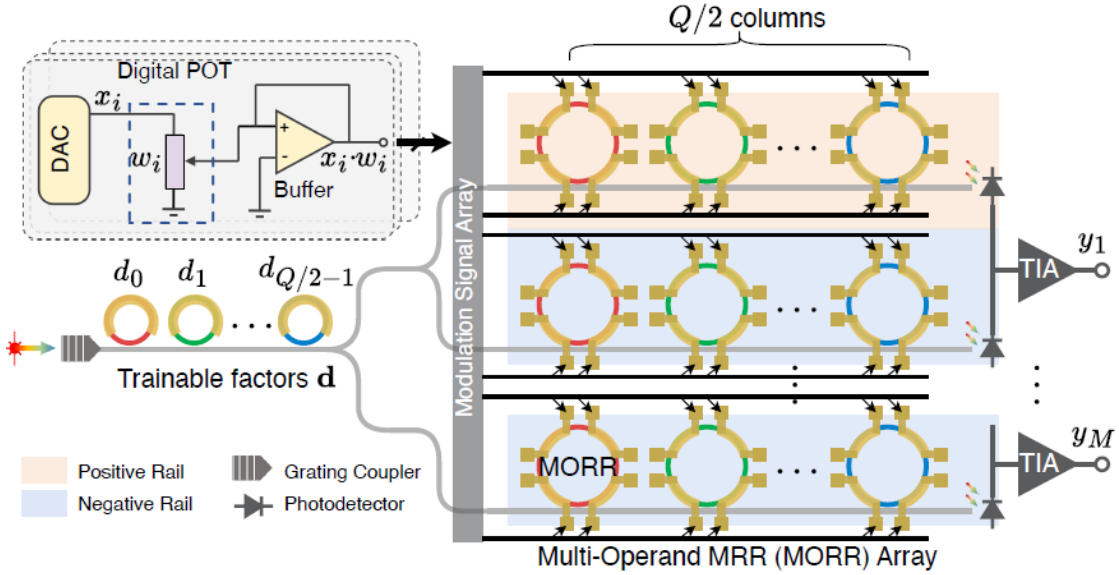
$$\phi_{rt} = \sum_{0}^{k-1} g_{\mathrm{mod}}\left(w_i \cdot x_i\right)$$

- Introduces ***on-chip nonlinearity*** by leveraging the intrinsic transmission characteristic of MRR.



$$T_t = f\left(\phi_{rt}\right) = \left|\frac{t_1 - t_2\alpha_{rt}e^{-j\phi_{rt}}}{1 - t_1 t_2 \alpha_{rt} e^{-j\phi_{rt}}}\right|^2 = f\left(\sum_{i=0}^{k} g_{\mathrm{mod}}\left(w_i \cdot x_i\right)\right)$$

MRR-based Multi-Operand Optical Neuron (M²OON)



Schematic of M²OON and proposed ONN architecture.

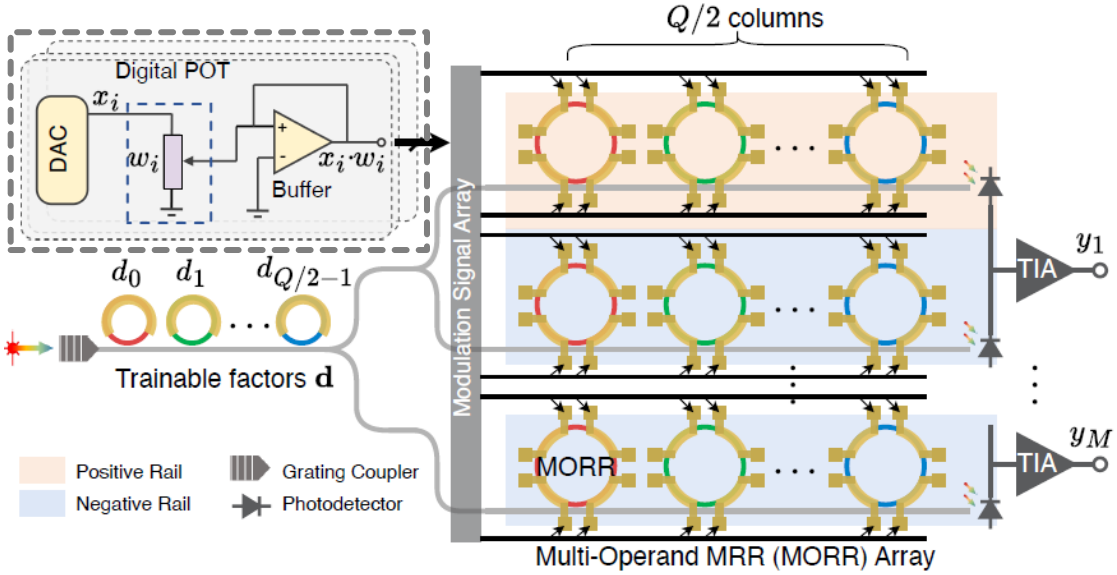
- $M \times N$ weight matrix is partitioned into $P \times Q$ blocks with a size of k
- To achieve full-range MVMs, the Q multi-operand MRRs for each row are split into **positive** and **negative** rails.
- Introduce a **learnable** balancing factor ***d*** to scale each MORR's output range adaptively, enhancing the expressivity of network.
- The output of each row can be expressed as

$$y_{m,+/-} = \sum_{q=0}^{Q/2-1} f_{\mathrm{MORR}} \left(\sum_{i=0}^{k-1} g_{\mathrm{mod}} \left(w_{mqi} \cdot x_{qi} \right) \right) d_q$$

Depend on modulation mechanism, e.g., thermal

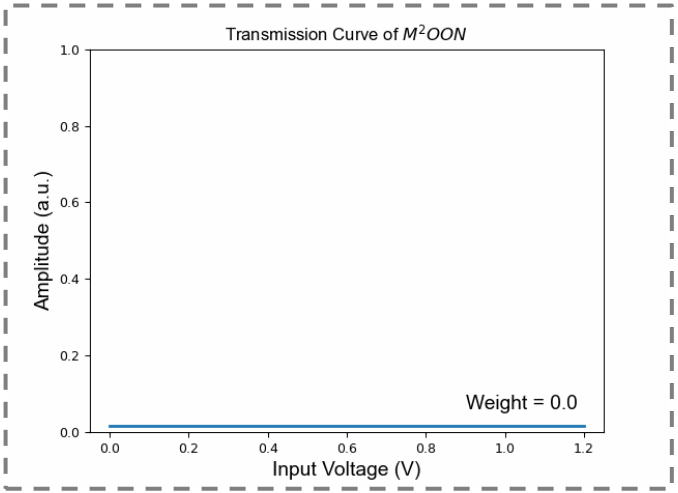
$d_q > 0$ for positive rail; $d_q < 0$ for negative rail

On-chip nonlinearity of an M²OON



Schematic of M²OON and proposed ONN architecture.

Response of ***Weights***

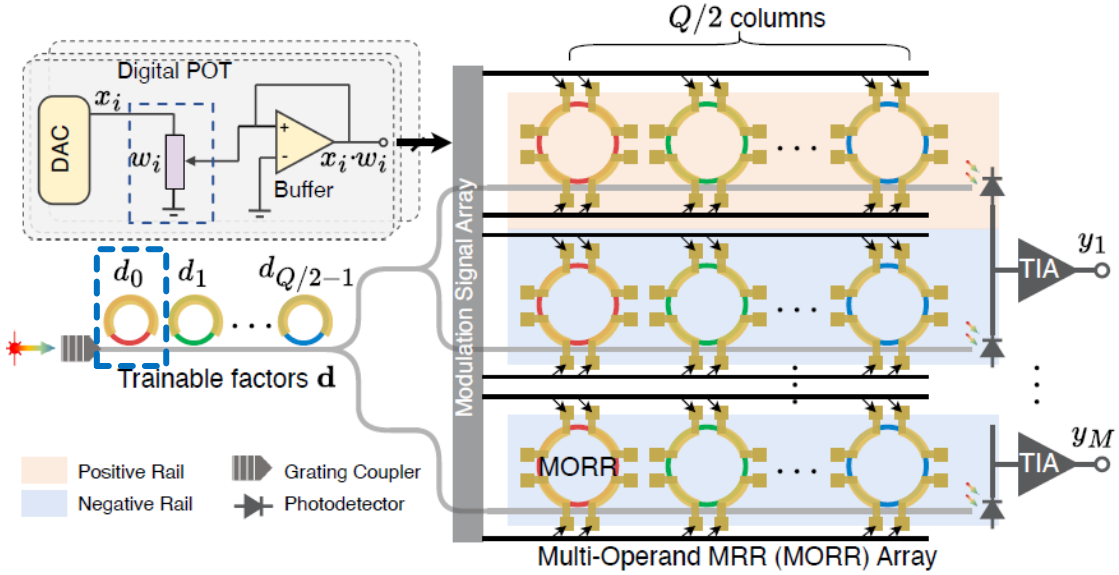


$$y_{m,+/-} = \sum_{q=0}^{Q/2-1} f_{\mathrm{MORR}} \left(\sum_{i=0}^{k-1} g_{\mathrm{mod}} \left(w_{mqi} \cdot x_{qi} \right) \right) d_q$$

w *are programed by digital potentiometers*

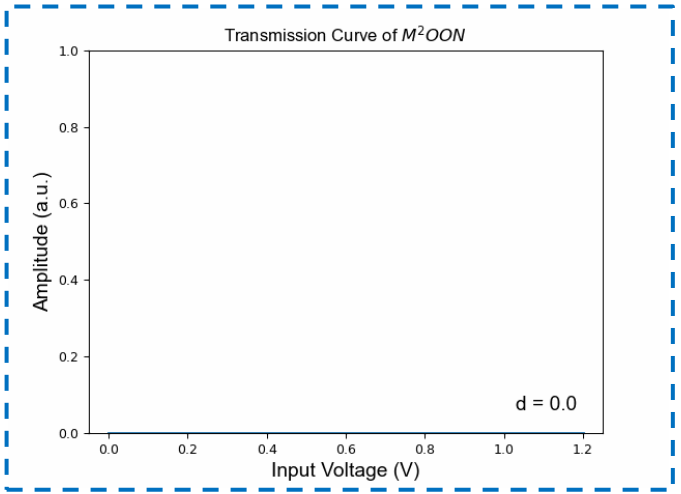
* All actuators are set to the same voltage for demonstration purposes

On-chip nonlinearity of an M²OON



Schematic of M²OON and proposed ONN architecture.

Tune Nonlinearity by ***d***

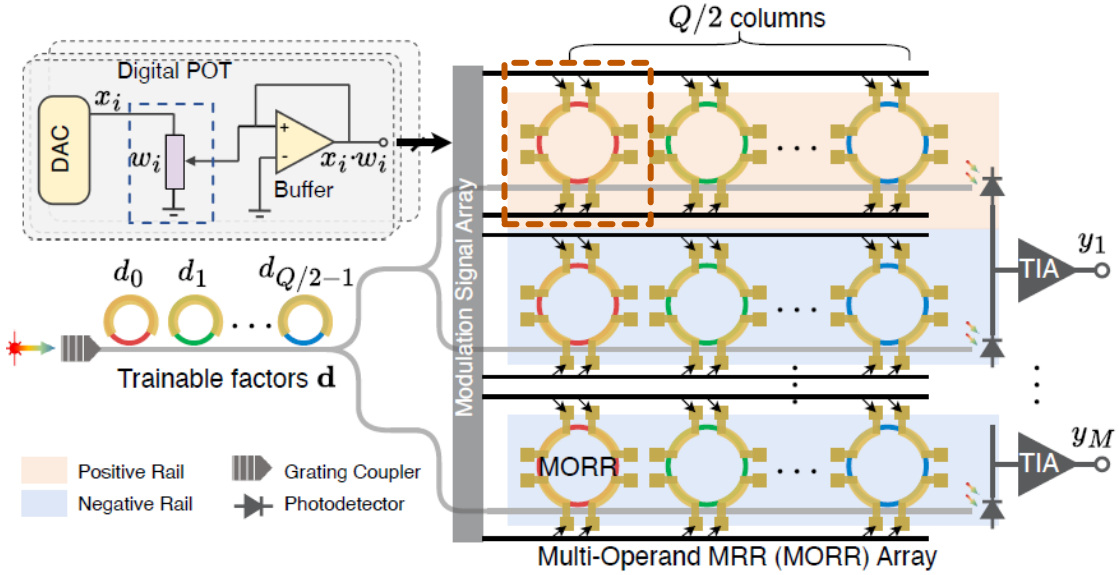


$$y_{m,+/-} = \sum_{q=0}^{Q/2-1} f_{\text{MORR}} \left(\sum_{i=0}^{k-1} g_{\text{mod}} \left(w_{mqi} \cdot x_{qi} \right) \right) d_q$$

d *are programed by regular MRRs*

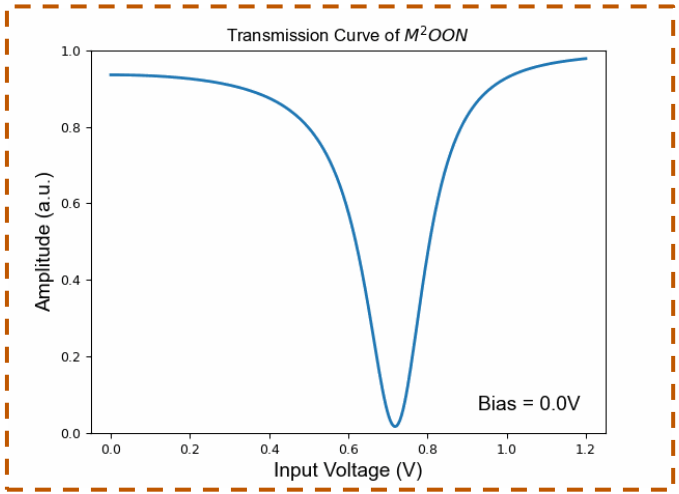
* All actuators are set to the same voltage for demonstration purposes

On-chip Nonlinearity of an M²OON

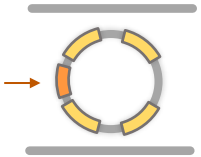


Schematic of M²OON and proposed ONN architecture.

Tune Nonlinearity by ***Bias***

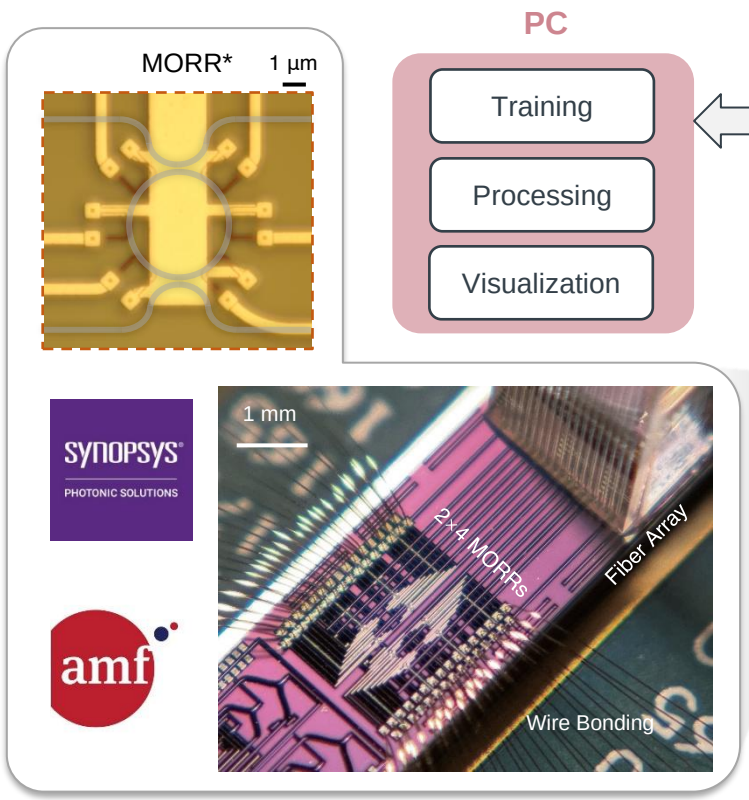


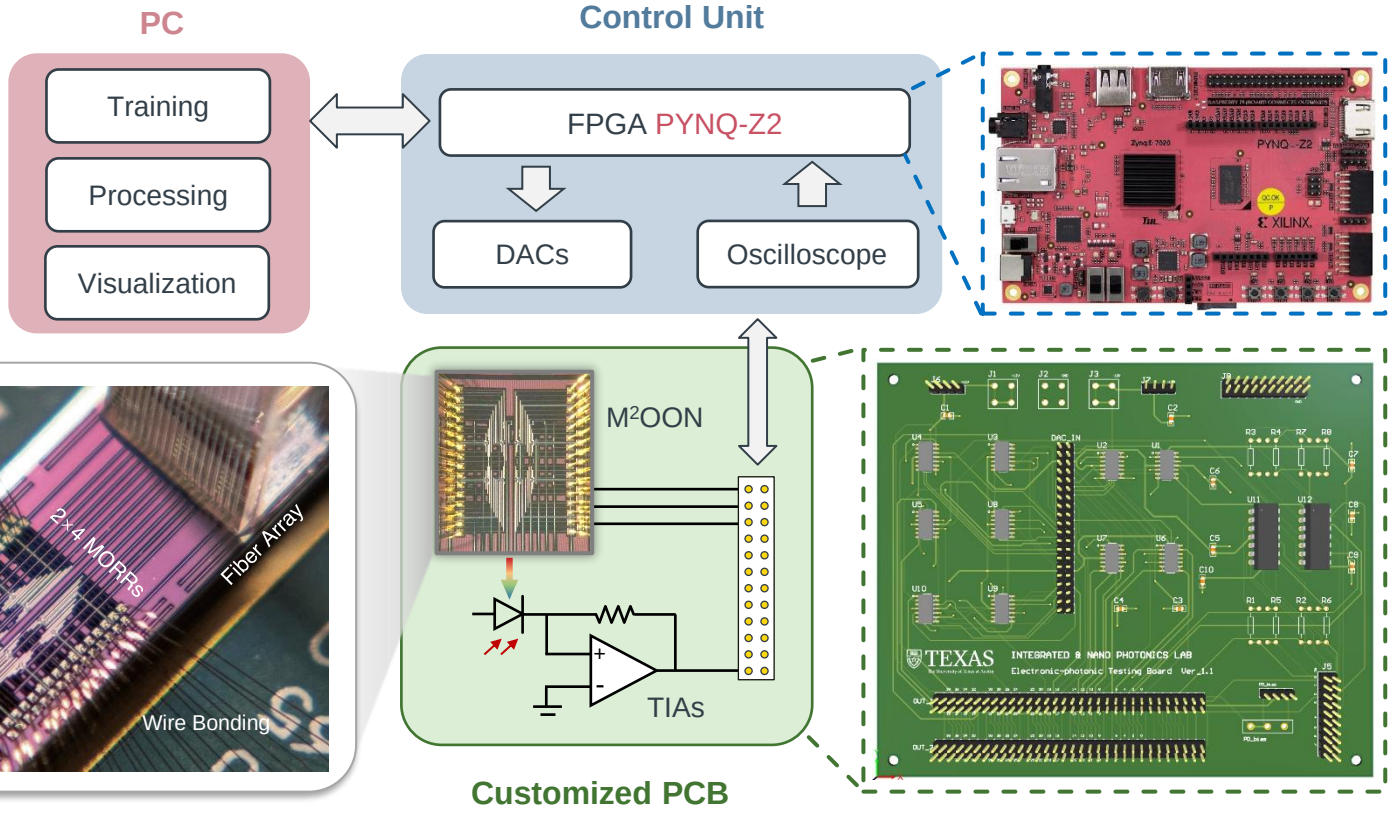
Bias *is applied to an additional* actuator, *besides the k* actuators *used for computation*



* All actuators are set to the same voltage for demonstration purposes

On-chip testing of a 4-operands M²OON





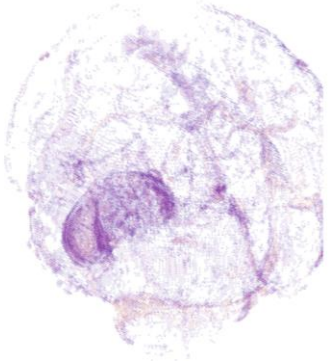
* 4 operands and 2 microheaters for calibration/tunable nonlinearity

Results: On-chip demonstration for nonlinear convolution process

- Input images: Human brain MRI image (3D) *from BraTS2020 Dataset*
- Convolve each slice with the 3 × 3 sobel kernel (on-resonate, *i.e.*, bias=0)
- For demonstration purposes, a threshold is set for 3D visualization
- Inherent tunable nonlinearity of M²OON allows for more **effective feature extraction**

On-chip Covn results

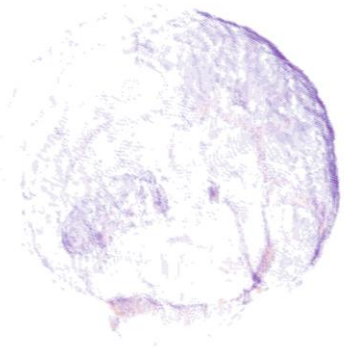
(output > 0.35)



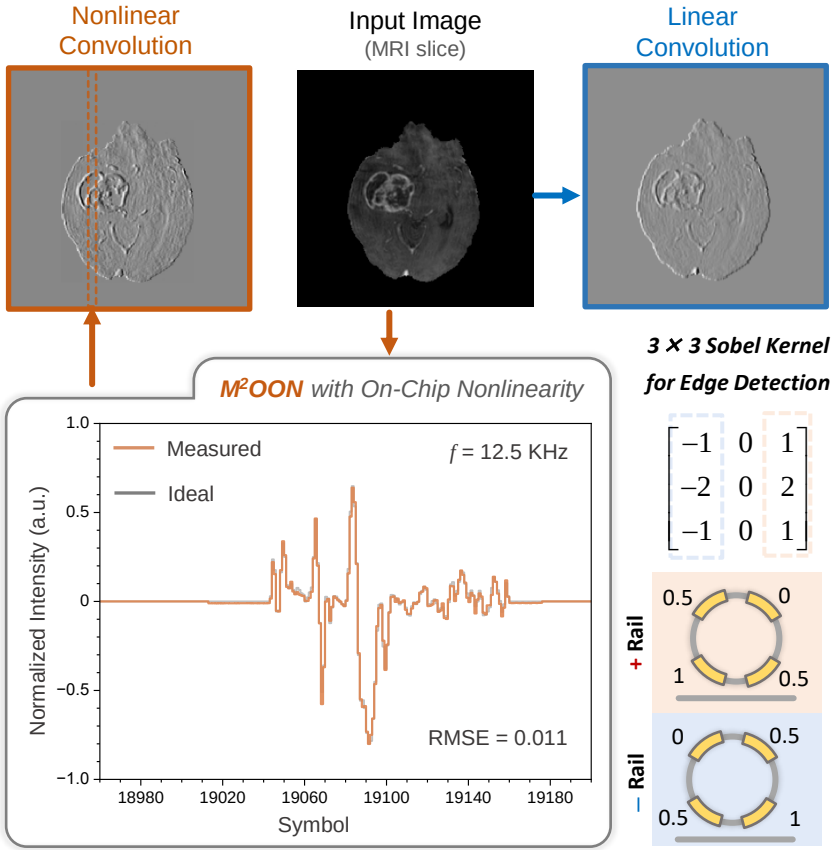
Tumor region is extracted

Linear Covn results

(output > 0.35)

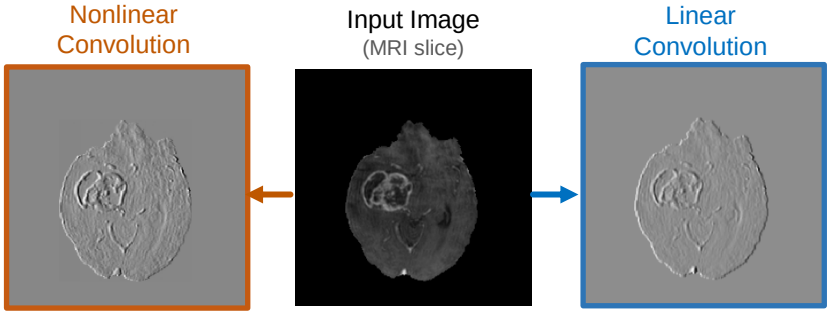


Tumor region is Vague



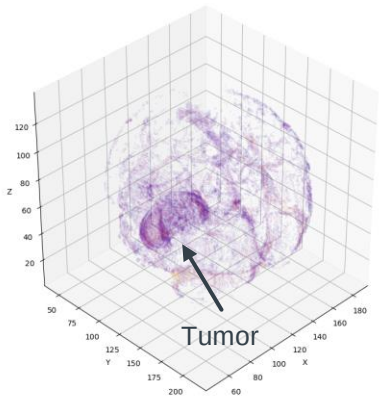
Results: On-chip demonstration for nonlinear convolution process

- Input images: Human brain MRI image (3D) *from BraTS2020 Dataset*
- Convolve each slice with the 3 × 3 sobel kernel (on-resonate, *i.e.*, bias=0)
- On-chip nonlinearity emphasize the boundary, increases contrast
- Inherent tunable nonlinearity of M²OON allows for more effective **feature extraction**



On-chip Covn results

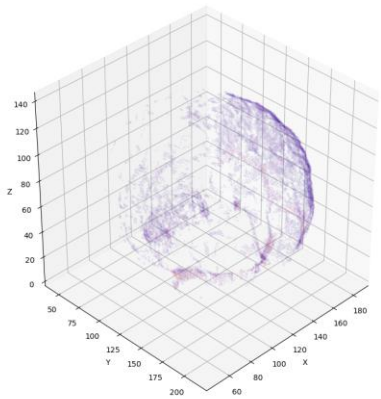
(output > 0.35)



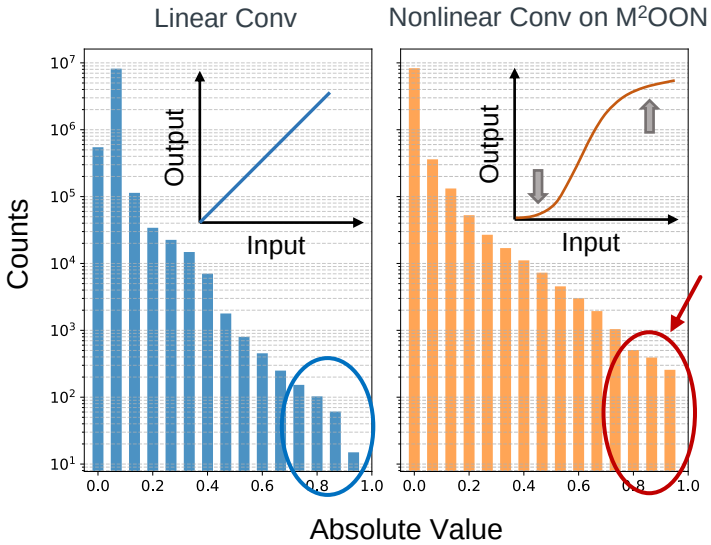
Tumor region is extracted

Linear Covn results

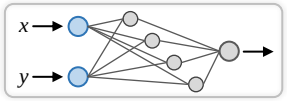
(output > 0.35)



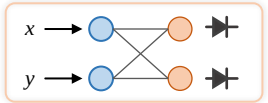
Tumor region is Vague



Results: On-chip demonstration for classification tasks



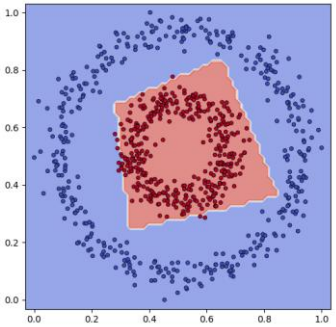
17 *trainable parameters for 2-layer NN*

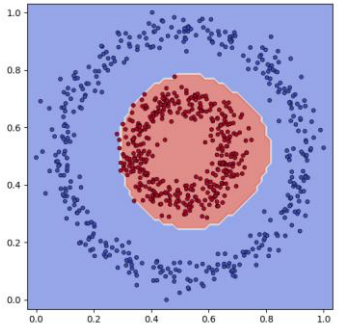


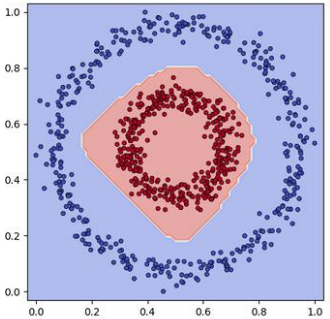
12 *trainable parameters for single-layer* ***ONN***

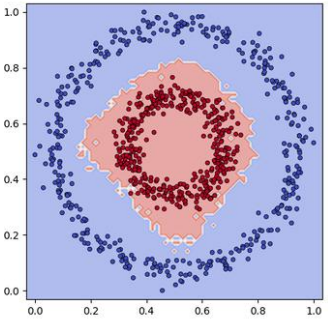
ReLU | Sigmoid | Simulation Result | Testing Result

Circle Distribution

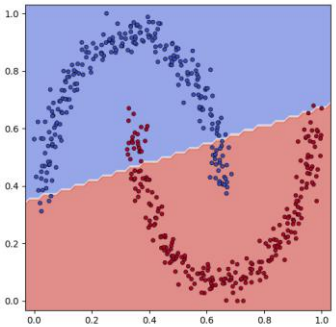


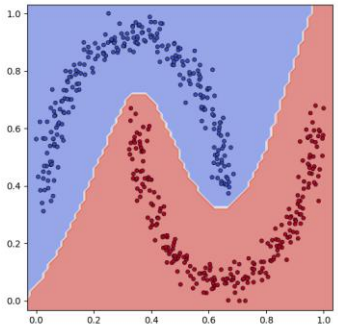


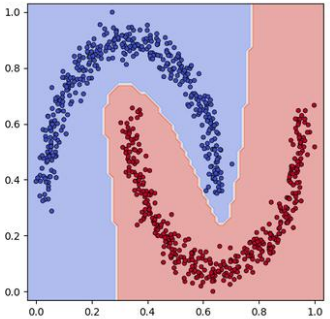


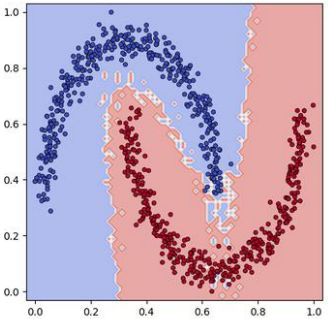


Hoof Distribution

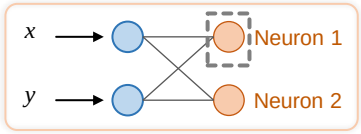


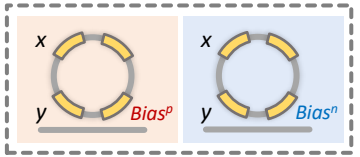




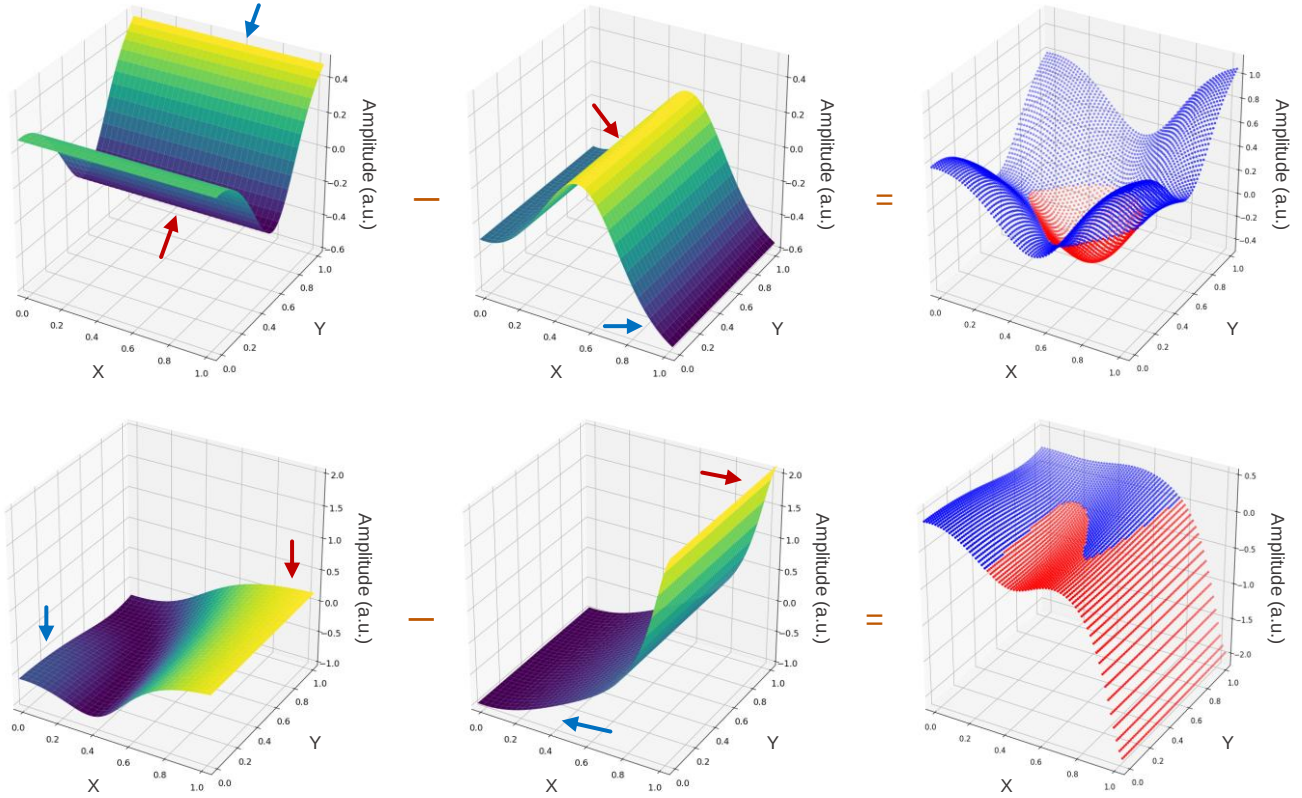


Results: Interpretability of M²OON-based ONN





Neuron 1 Neuron 2



Results: Image recognition with block-circulant optical neuron

Dataset: PTB-XL electrocardiography (ECG)

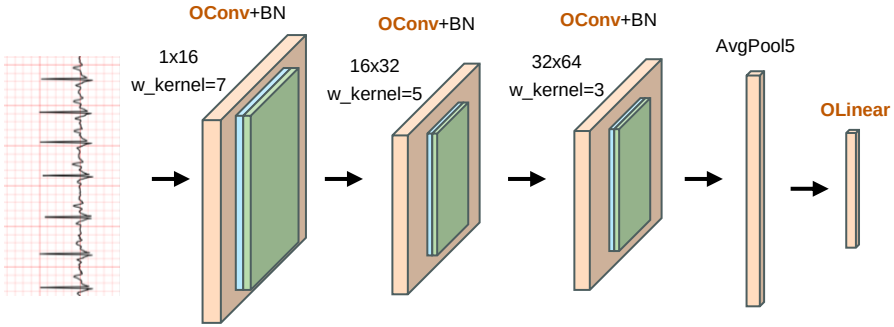
Structure: 3-layer CNN

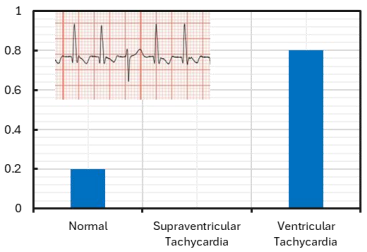
Results:

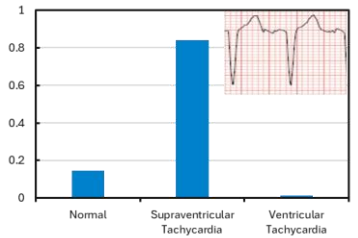
- ~95% measured accuracy for arrhythmia detection

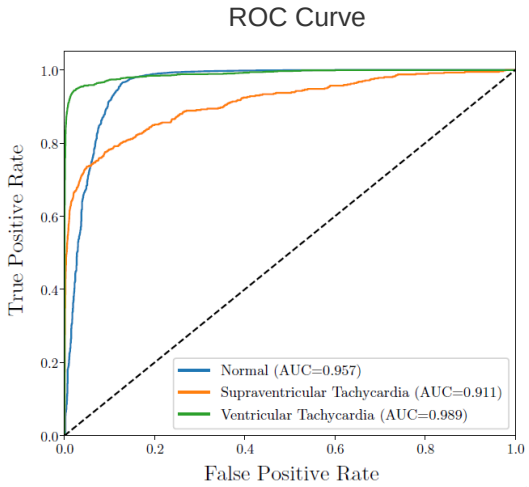
- 6-bit control precision for input encoding
- 8-bit control precision for weight encoding

More complicated datasets and analysis will be discussed in journal version









Summary

- We present a hardware-efficient ONN based on multi-operand MRR to implement tensor operations.
  - — Squeeze tensor operation on a single device
  - — on-chip tunable nonlinearity for higher representability of neural network
  - — >95% measured accuracy on arrhythmia detection task
- An FPGA-based ONN testing platform and AI-assisted hardware-aware training framework are developed.
- Experimentally demonstrated optical convolution process with emphasized features
- More details and complex tasks will be discussed in journal publication

Thank you !

Any Question?

Acknowledgment

Support from AFRL/AFOSR MURI project

Supervisors: Prof Ray. T. Chen and prof David Z. Pan

Colleagues: Jiaqi Gu, Chenghao Feng， Hanqing Zhu， Rongxing Tang