Light for AI: Hardware-Software Codesign in Optical Neural Networks

Jiaqi Gu, Chenghao Feng, Ray. T. Chen, and
David Z. Pan
Dept. of Electrical and Computer Engineering
The University of Texas at Austin

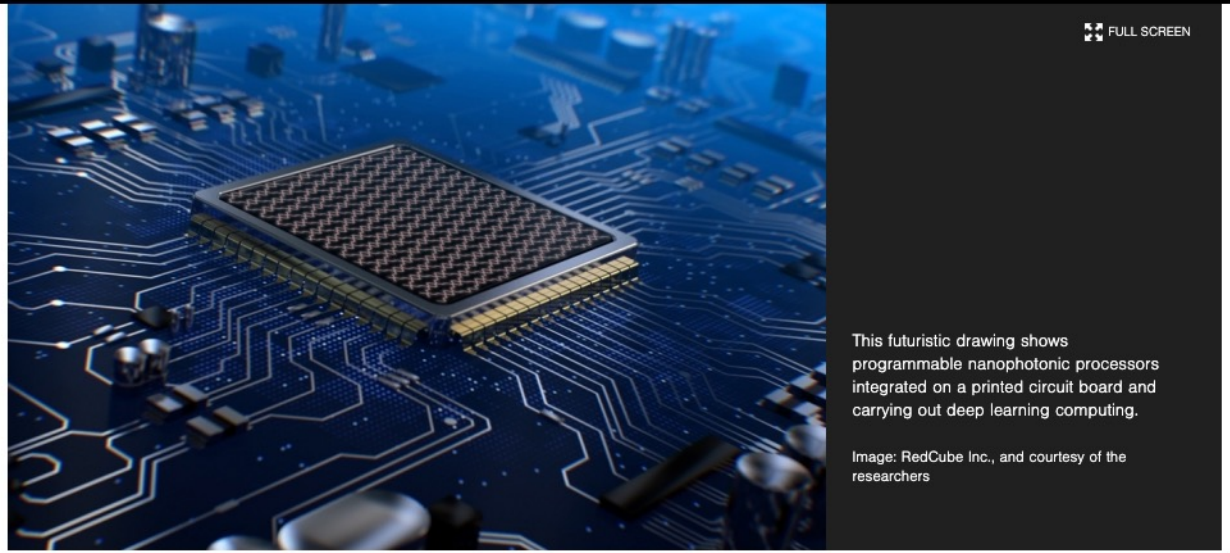
This work is supported in part by MURI and ONR

Photonic AI Chip

Based on optics/photonics
➔ photonic ICs

MIT News
ON CAMPUS AND AROUND THE WORLD

Browse or Search



This futuristic drawing shows programmable nanophotonic processors integrated on a printed circuit board and carrying out deep learning computing.

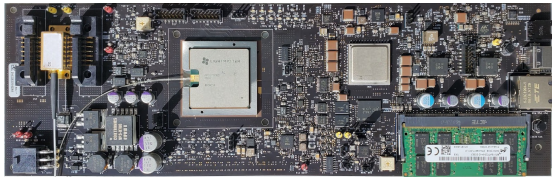
Image: RedCube Inc., and courtesy of the researchers

New system allows optical “deep learning”

Neural networks could be implemented more quickly using new photonic technology.

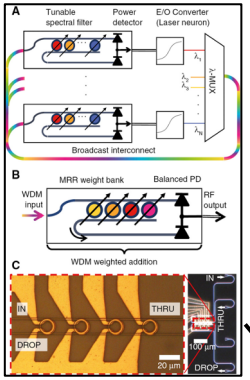
LIGHTELLIGENCE

LIGHTMATTER

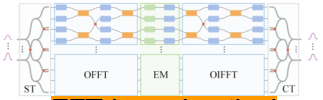


VC-backed startups from MIT

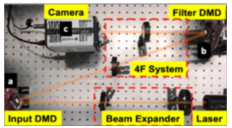
ONN Progress



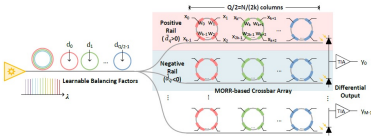
MRR Neural Network
[Brunner+, 2016]
[Tait+, SciRep 2017]
Princeton



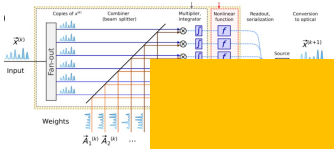
FFT-based optical neural network
[Gu+, ASPDAC2020]
UT

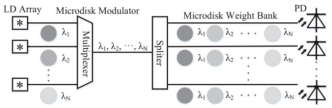


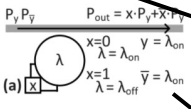
Free-space ONN
[Miscuglio+, Optica2020]
GWU

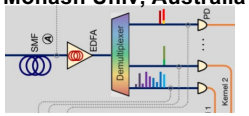


MORR ONN
[Gu+, DATE2021]
UT



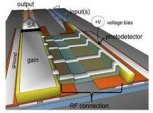


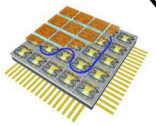




WDM Comb
[Xu+, Nature2021]
Monash Univ, Australia

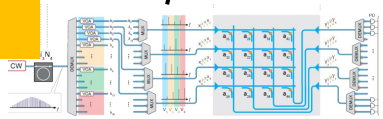
PCM Xbar
[Miscuglio+, APR2020]
GWU



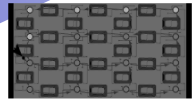


Optical Spike Neural Network
[Tait+, 2016]
Princeton

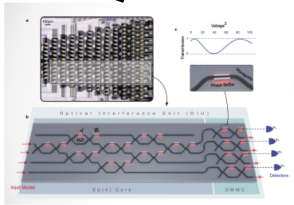
Spiking ONN: PCM
[Feldmann+, Nature 2019]
Munster, Oxford



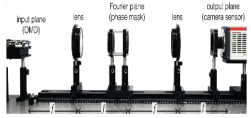
PCM Xbar
[Feldmann+, Nature2021]
Munster, Oxford



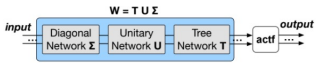
Optical Reservoir Computing
[Vandoorne+, NatureComm 2014]
Ghent University



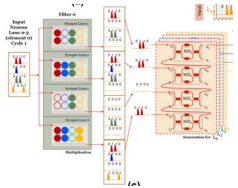
MZI-based Neural Network
[Shen+, Nature Photonics 2017]
MIT



Free-space ONN
[Chang+, SciRep 2018]
Stanford



SlimmedONN
[Zhao+, ASPDAC2019]
UT



PIXEL, MZI Multiplier
[Shieflett+, HPCA2020]
Ohio Univ

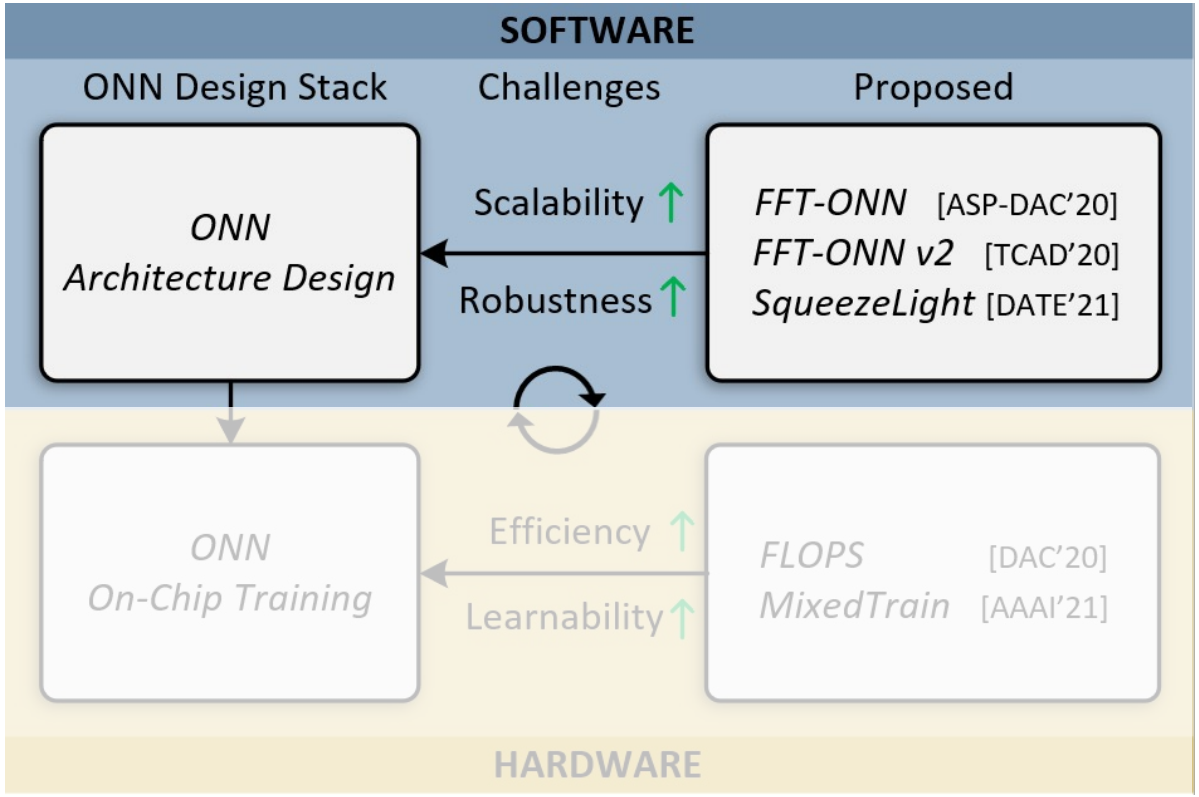
Area Cost
Robustness
Learnability

2016
2017
2020
2021

3

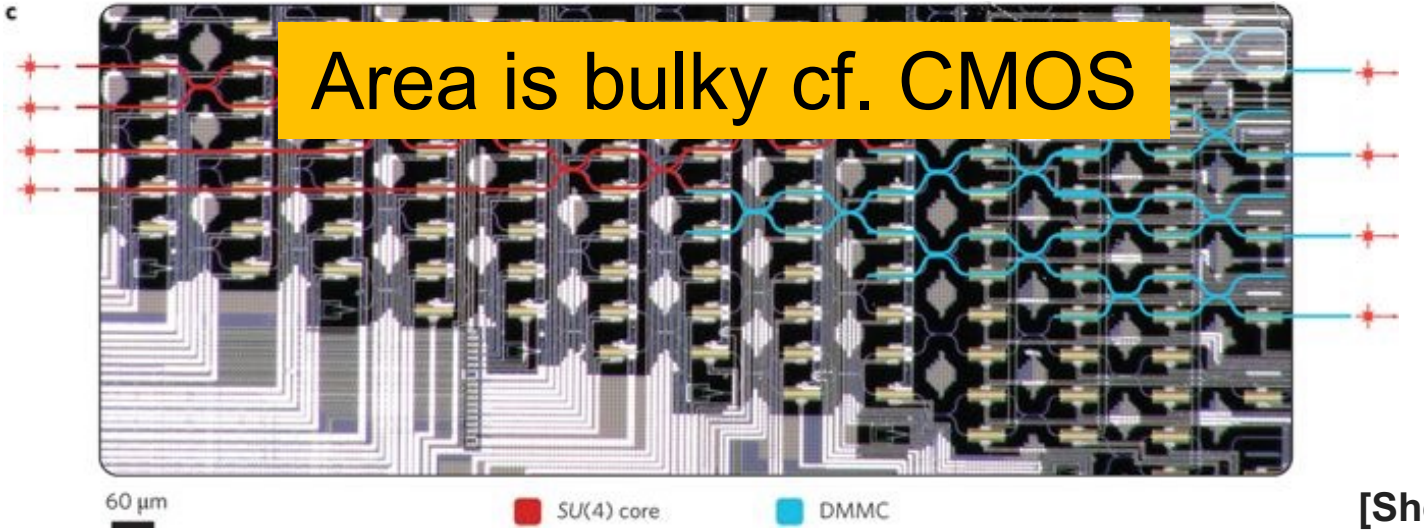
Outline

- Hardware-Efficient ONN Architecture
- Efficient On-Chip Learning for ONNs



Optical Neural Network (ONN) Processor

- Light in and light out (analog computing)
  - Speed-of-light floating point **matrix-vector multiplication**
  - >100GHz detection rate
  - Ultra-low energy consumption when configured

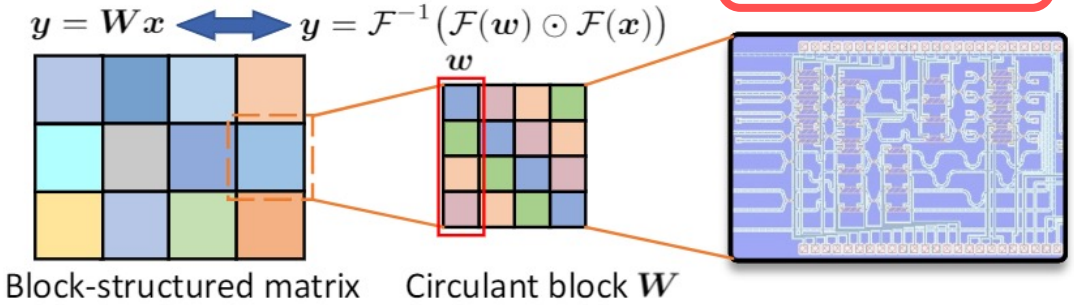


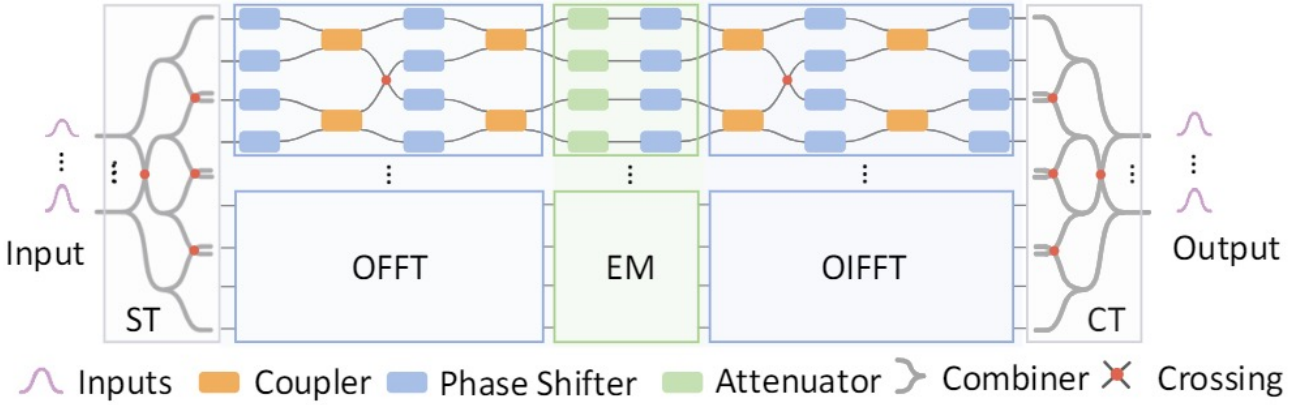
[Shen, Harris, et al., *Nature Photonics* 2017]

Our FFT-based ONN [Gu+, ASP-DAC'20, BPA]

- Efficient circulant matrix multiplication in Fourier domain
- 2.2~3.7× area reduction, no accuracy loss

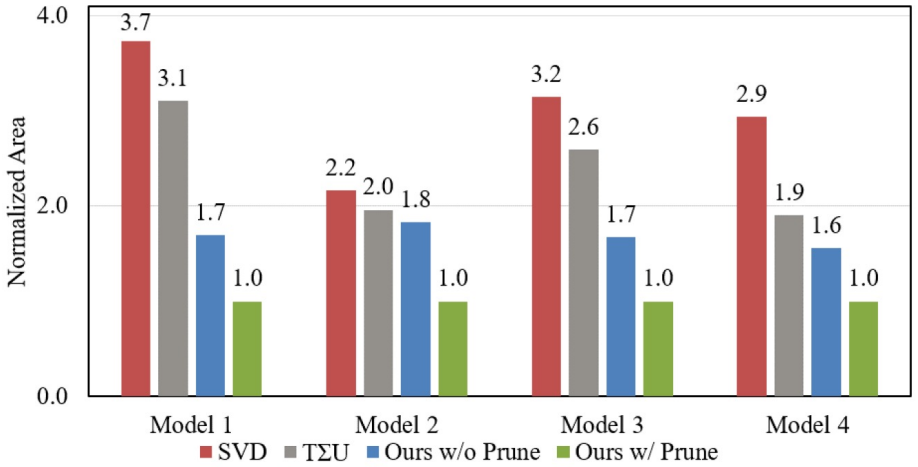
$$O(m^2 + n^2) \rightarrow O\left(\frac{mn}{k}\log_2 k\right)$$





Our FFT-based ONN [Gu+, ASP-DAC'20, BPA]

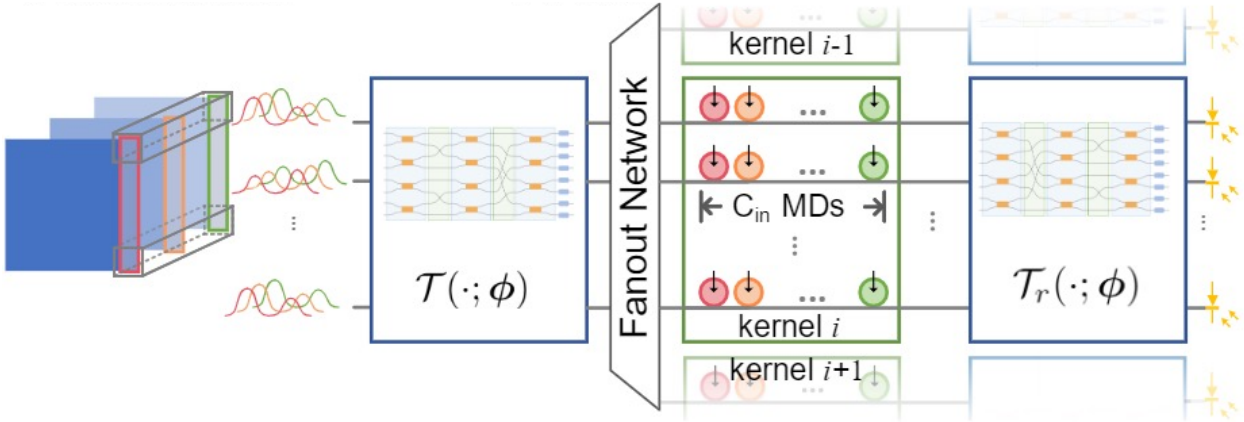
- Advantages
  - 2.2~3.7× smaller than MZI-ONN
  - Without accuracy loss
  - More robust & lower latency
    - Lower network depth

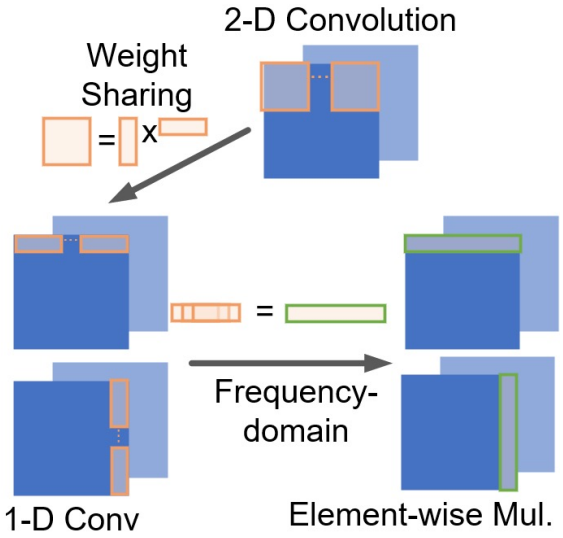


- Remaining issues
  - **CNN**: not support efficient convolutional neural networks (CNNs)
  - **Power**: no significant power improvement in device tuning power
  - **Expressivity**: not fully-explore the learnability of the structure with fixed OFFT

Proposed Frequency-Domain ONN [Gu+, TCAD'20]

- From fixed FFT to general frequency-domain transform (FFT-ONN v2)
- From real spatial matrix to complex frequency-domain matrix
- WDM-based highly parallel CNN
- More learning expressiveness and optimization flexibility

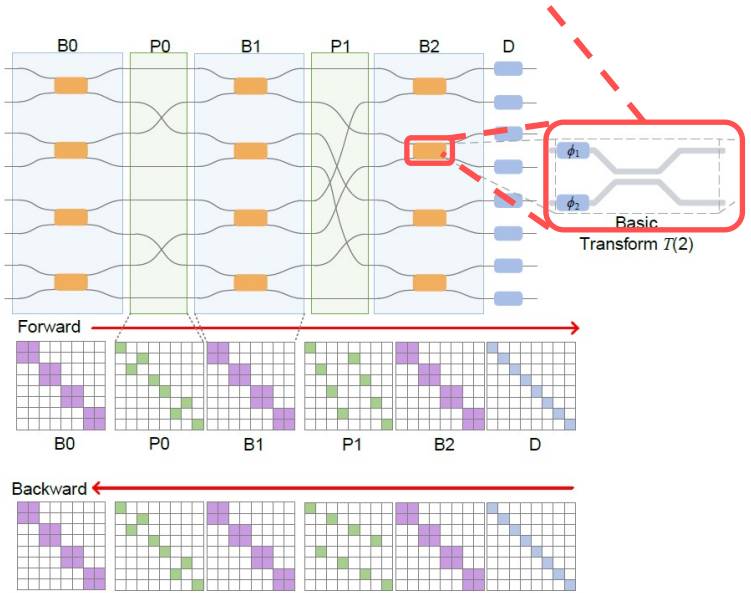




Learnable Frequency-Domain Transform

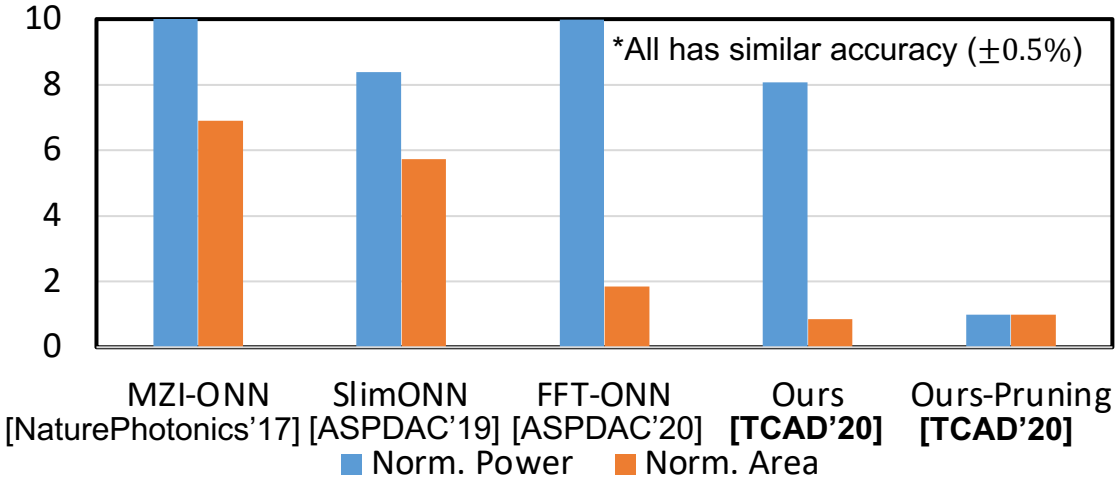
- Not necessary to use Fourier transform $\mathcal{T}$
  - Fixed topology + trainable unitary block
- Not necessary to use inverse transform $\mathcal{T}^{-1}$
  - Arbitrary original and reversed transforms
- Automatically learn the best transform pair
- Sparsify the phase shifters with pruning
  - Smaller footprint
  - Lower power
  - Less noise

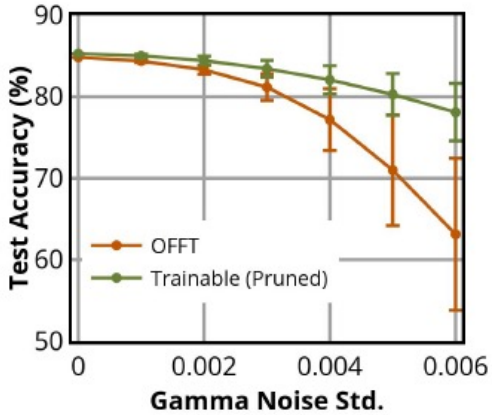
$$\mathcal{T}(2) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix} \begin{pmatrix} e^{j\phi_1} & 0 \\ 0 & e^{j\phi_2} \end{pmatrix}$$
$$= \frac{1}{\sqrt{2}} \begin{pmatrix} \cos\phi_1 + j\sin\phi_1 & -\sin\phi_1 + j\cos\phi_1 \\ -\sin\phi_2 + j\cos\phi_2 & \cos\phi_2 + j\sin\phi_2 \end{pmatrix}$$



Experimental Results

- MLPs: with learnable transforms and pruning
  - 2.2× smaller area than FFT-ONN; ~7× smaller area than MZI-ONN
  - >10× lower phase shifter programming power than MZI-ONN & FFT-ONN
  - Much better robustness under phase noises
- CNNs: with learnable transforms and pruning
  - 5.6-11.6× smaller than MZI-ONN

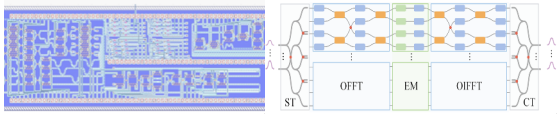




Area Bound of Optical Neural Networks

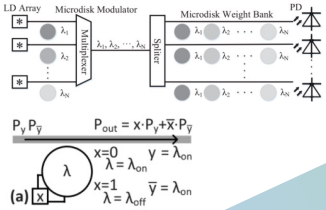
- Compact design is in demand
- Area is bounded by 1 MAC/MRR

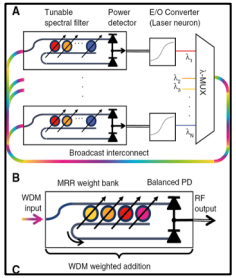
How to break through the area bound ?



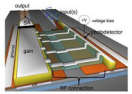
FFT-based ONN
[Gu+, ASPDAC2020]
[Gu+, TCAD2020]

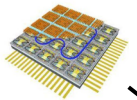
Holylight and Lightbulb: MRR&PCM
[Liu+, Zokaee+ DATE'2019, 2020]



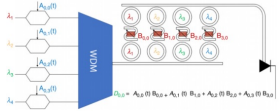


MRR Neural Network
[Brunner+, 2016]
[Tait+, SciRep 2017]

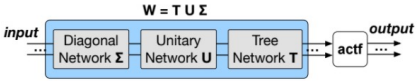




Optical Spike NN
[Tait+, 2016]

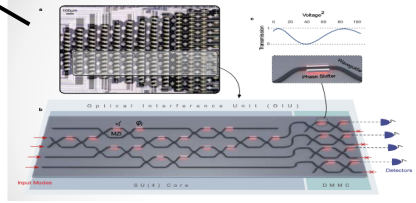


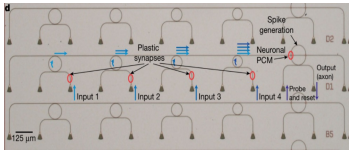
PCM Xbar
[Miscuglio+, APR2020]



SlimmedONN
[Zhao+, ASPDAC2019]

MZI-based Neural Network
[Shen+, Nature Photonics 2017]





Spiking ONN: PCM
[Feldmann+, Nature 2019]

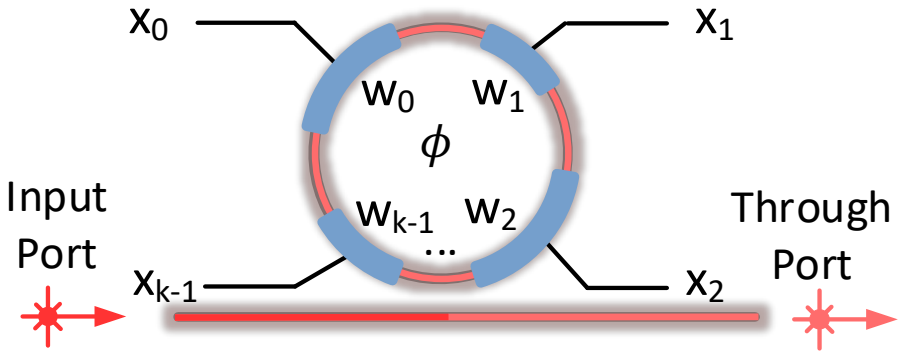
2016 2017 2018 2019 2020

Area Cost Still a Concern

Squeezing with Multi-Operand Ring [Gu+, DATE'21]

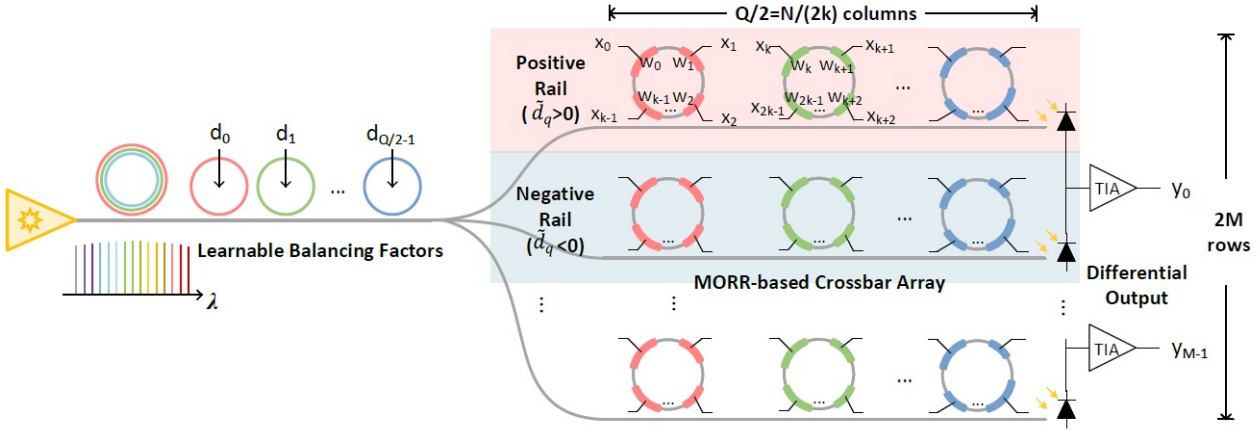
- MORR: $\boldsymbol{k}$-segment controllers on one micro-ring
- Single-device vector dot-product

  Round-trip phase: $\phi \propto \sum_{i=0}^{k-1} w_i x_i^2$



MORR-based ONN: SqueezeLight [Gu+, DATE'21]

- MORR array to perform matrix multiplication + nonlinear activation
- Differential rails support positive/negative neurons



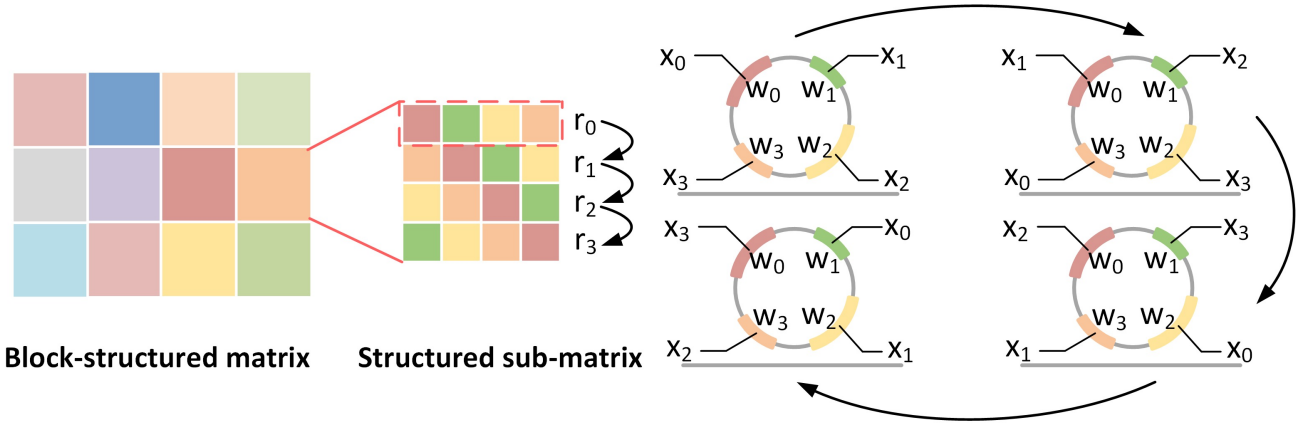
- Learnable balancing factors d_i
  - Adaptive MORR output range
  - Enhanced expressivity

$k^2\times$ fewer devices
$k^2\times$ area reduction
$2k\times$ fewer wavelengths

MORR Sharing

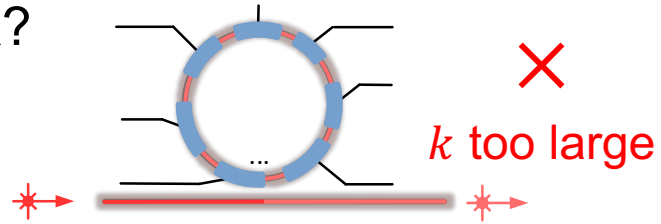
[Gu+, DATE'21]

- Squeeze a $k \times k$ block into one MORR
  - Share weights in multiple rows → share the same MORR

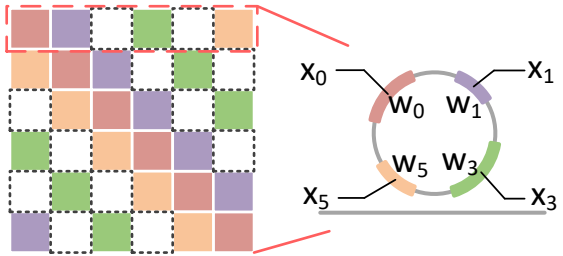


Sparsity Exploration - Pruning

- How to squeeze larger block into one MORR?
  - #Operand limit on one MORR
  - Manufacturing, crosstalk, ...



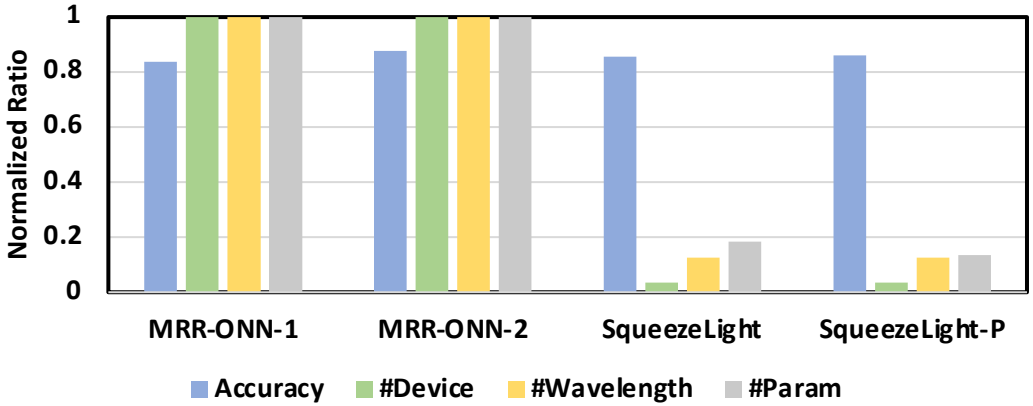
- Sparsify blocks via fine-grained structured pruning
  - 4-op MORR ←→ 6×6 block (33% sparsity)
  - 4-op MORR ←→ 8×8 block (50% sparsity)
  - Support larger blocks with small MORR
  - Pruning-aware training

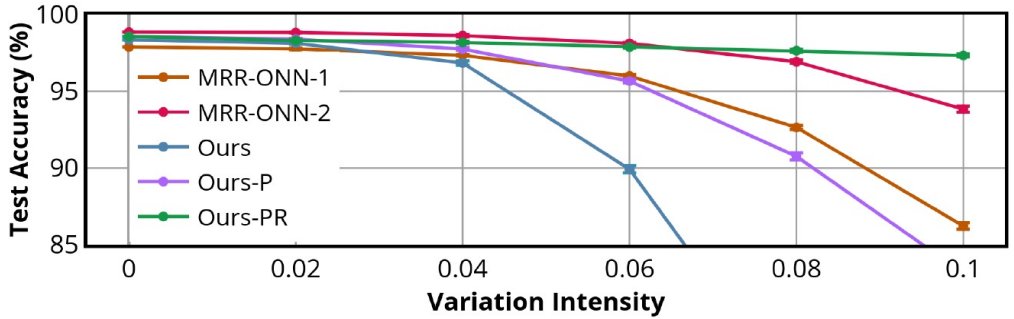


Sparse structured sub-matrix

Comparison: Accuracy, Scalability, Robustness

- Compare with SoTA MRR-ONNs on MNIST, FMNIST, CIFAR-10
  - All-pass MRR-ONN-1
  - Add-drop MRR-ONN-2
- Comparable expressivity
- 23×-32× less device usage
- 8× fewer wavelength usage
- >5× fewer parameters
  - 50% sparsity
  - No accuracy drop
- Better noise-robustness

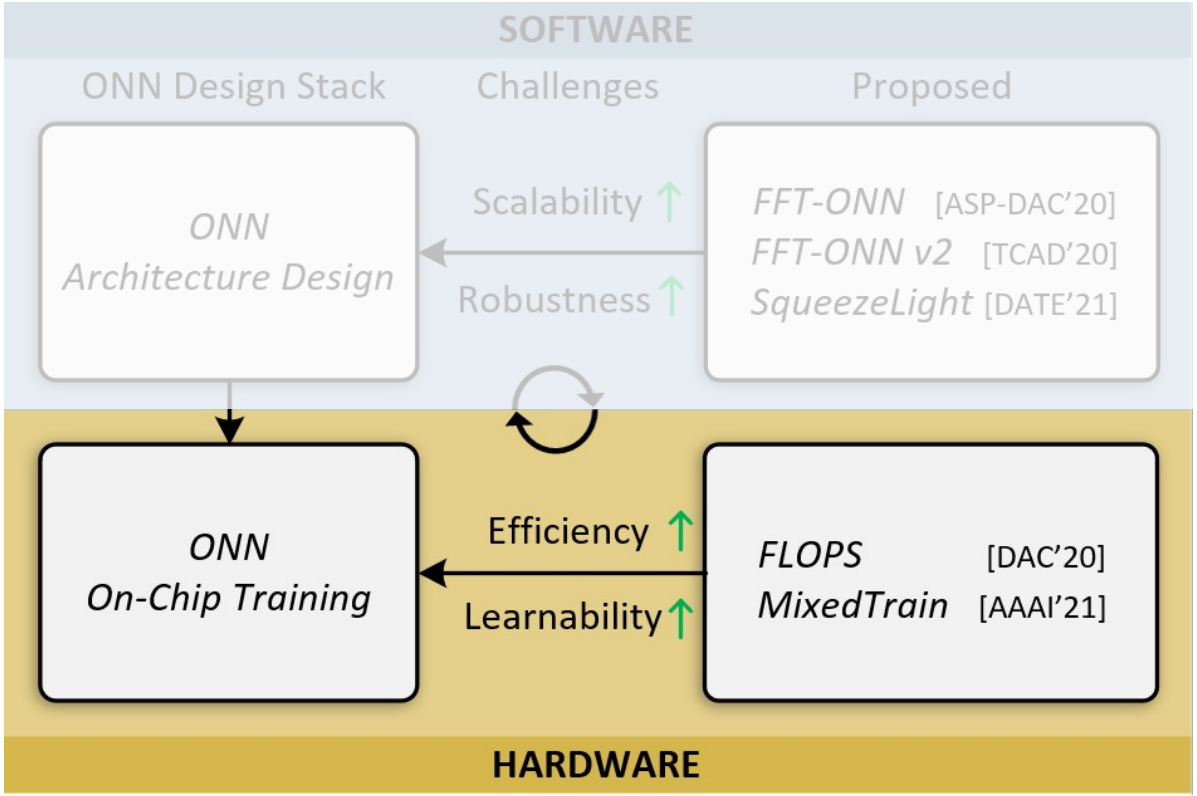




MRR-ONN-1 [Liu+, DATE'2019]
MRR-ONN-2 [Tait+, SciRep 2017]

Outline

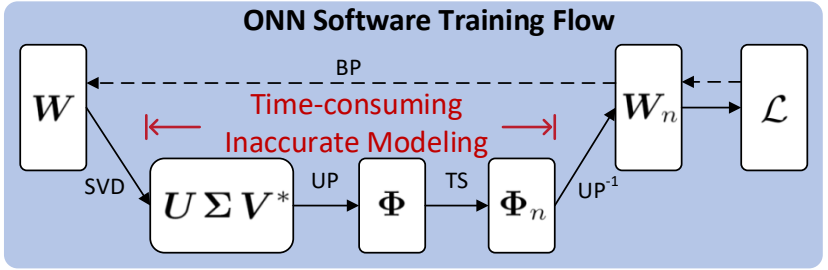
- Hardware-Efficient ONN Architecture
- Efficient On-Chip Learning for ONNs



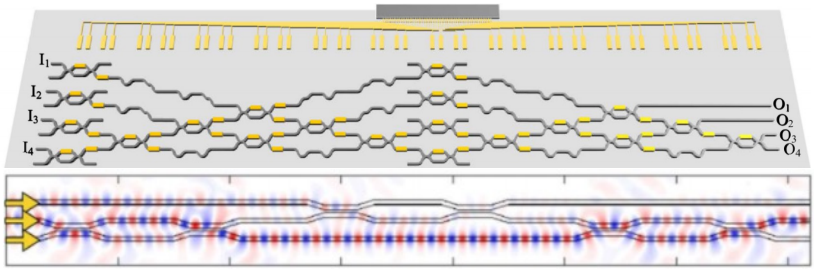
ONN On-Chip Training

- What is ONN on-chip (on-device) training
  - *In-situ* **accuracy recovery** on **non-ideal** photonic integrated circuits

- Why on-chip training
  - Inaccurate software modeling
  - Limited speed: ~1*s* per iter



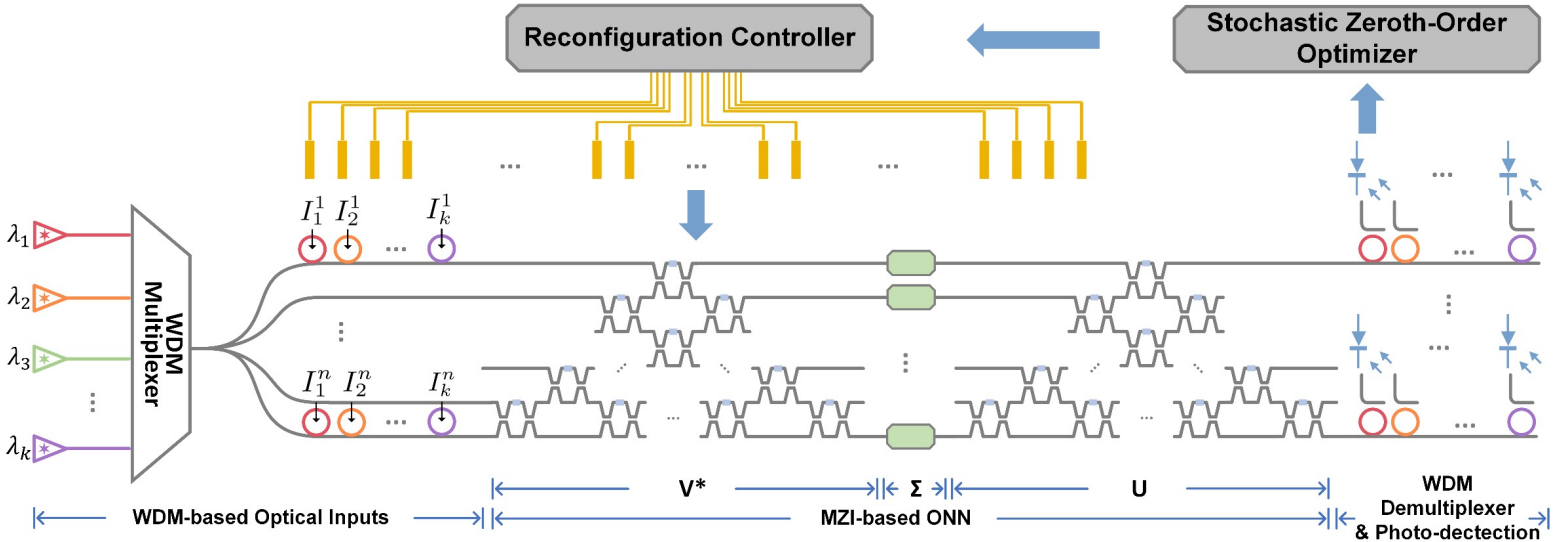
- Prior on-chip training protocols
  - [Hughes+, *Optica*'18] [Zhang+ *OE*'19] [Zhou+, *JSTQE*'19]
  - ~100x faster: ~1 *ms* per iter
  - Unscalable: <100 MZIs
  - Divergence issues
  - Limited efficiency



[Zhou+, *JSTQE*'19] [Hughes+, *Optica*'18]

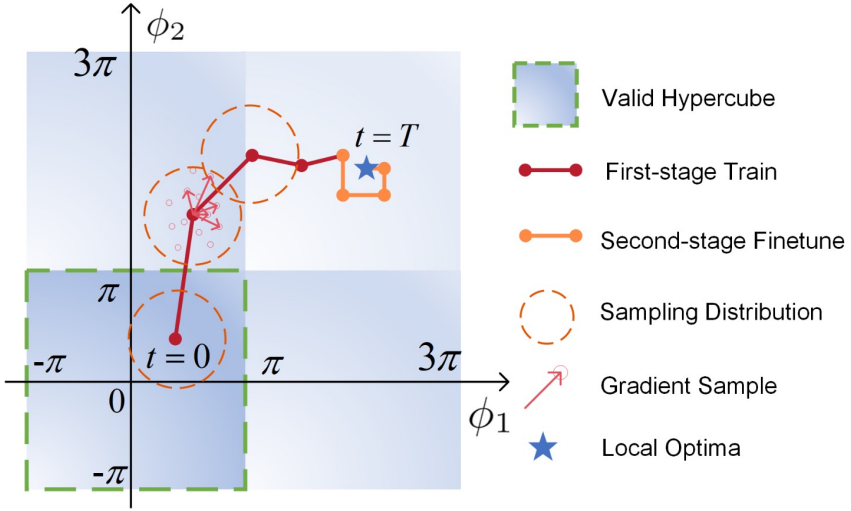
Our Method: FLOPS [Gu+, DAC'20 BPC] [NSF Workshop'20, BPA]

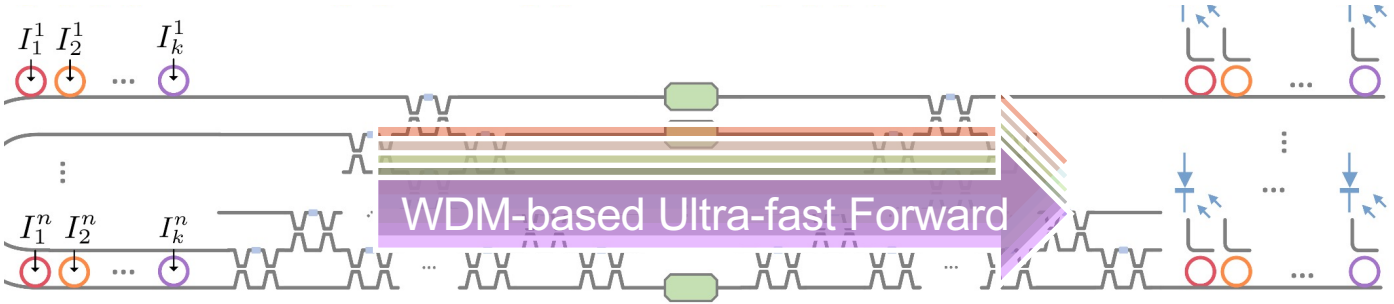
- ONN on-chip learning via stochastic zeroth-order optimization
  - **Efficiency**: WDM-based parallel gradient estimation
  - **Accuracy**: Two-stage learning protocol with high accuracy
  - **Robustness**: Robust learning under *in situ* device variations



FLOPS and FLOPS+

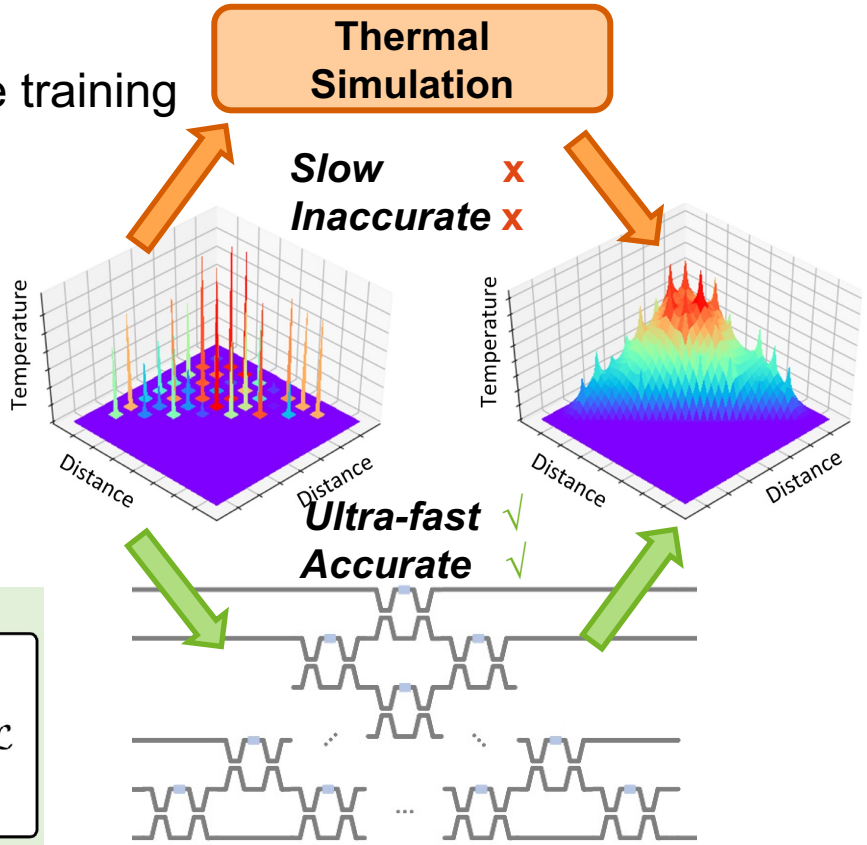
- FLOPS
  - Fast exploration with ZO gradient
  - High parallelism with WDM
  - Faster convergence
  - May not be accurate enough
- FLOPS+ with ***SparseTune***
  - Sparse coordinate-wise fine-tuning
  - Improve accuracy via searching
  - Better convergence

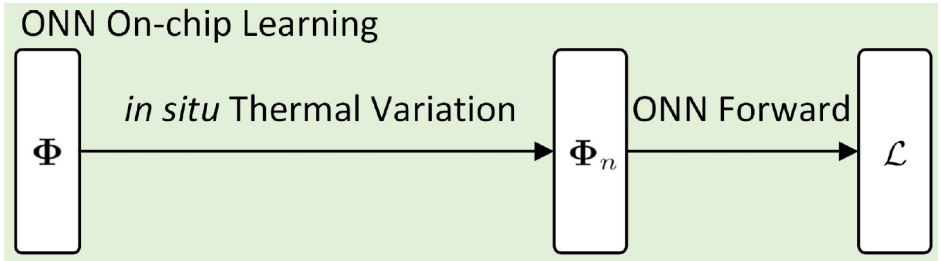




Robust On-Chip Learning

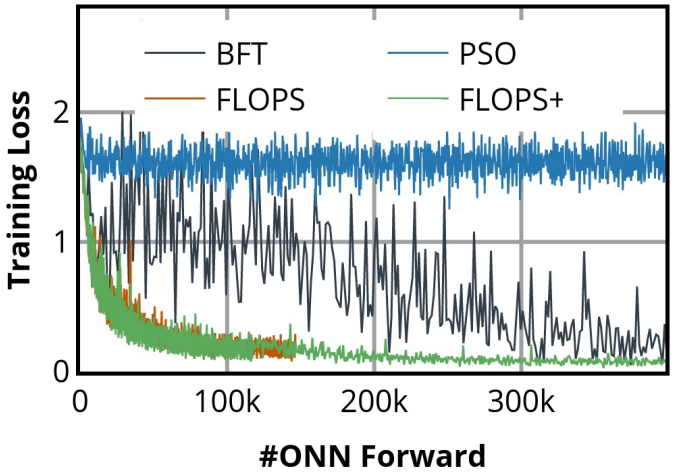
- Thermal crosstalk variations
  - Typically not considered in software training
  - Time-consuming
  - Inaccurate
- Built-in robustness handling on-chip
  - Ultra-fast: ~1 μs
  - Accurate: physical noise model





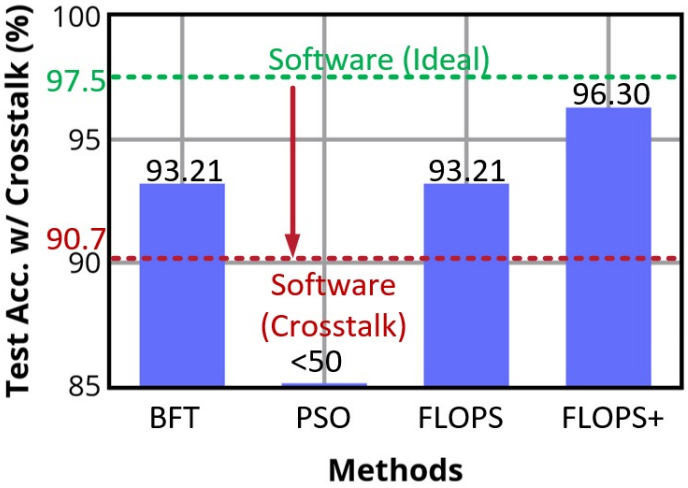
Experimental Results (Efficiency & Robustness)

- Efficient learning on vowel recognition task **[Deterding+, 1989]**
  - FLOPS: **4x** more query-efficient (800 k vs. 200 k)
  - FLOPS+: **2x** more query-efficient (800 k vs. 400 k)
  - **3%** more robust than previous on-chip training approaches



ONN config: 10-24-24-6 (960 MZIs)

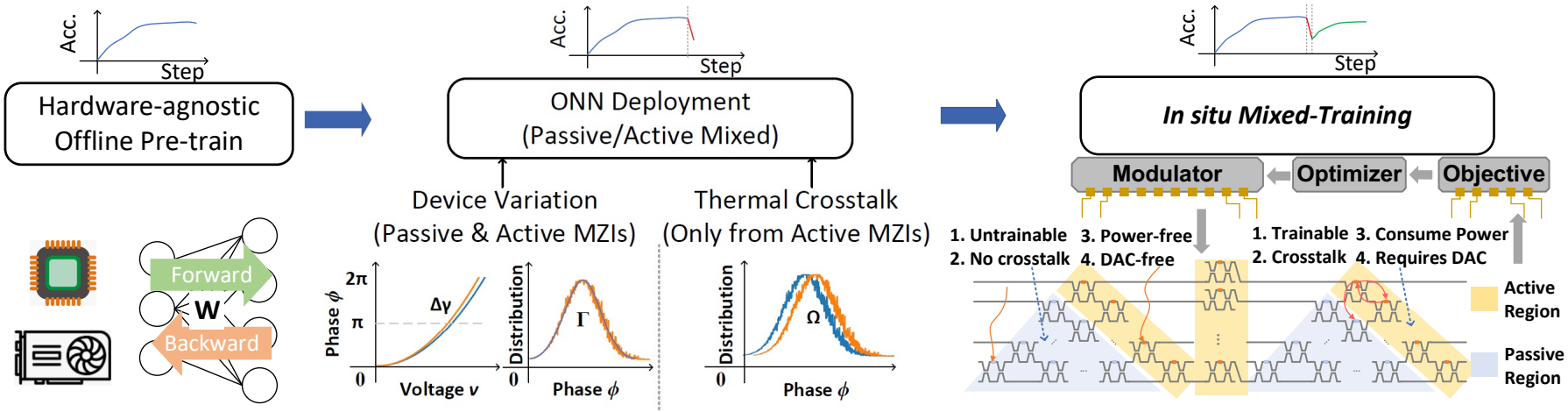
[Zhou+, 2019]
[Zhang+, 2019]



ONN config: 10-24-24-6 (960 MZIs)

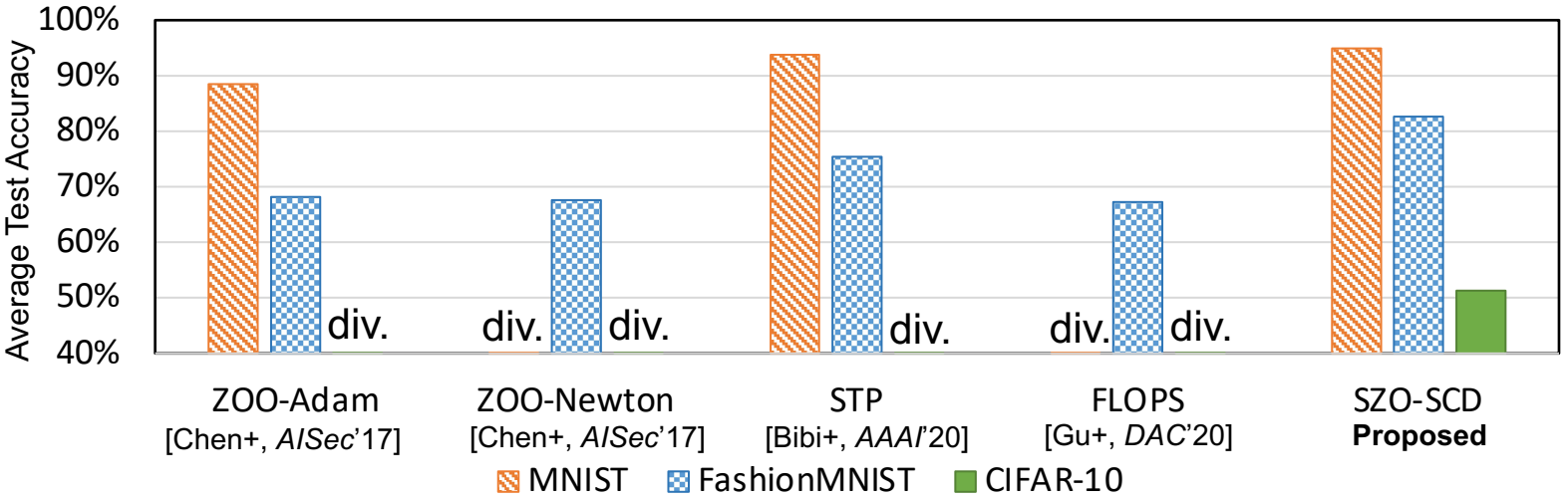
Enhanced Solution: MixedTrain [Gu+, AAAI'21]

- Initialization: Hardware-agnostic offline pretraining
- Deployment: Mixed active/passive regions + non-ideality
- In-situ Learning: Mixed training/sparsity strategies



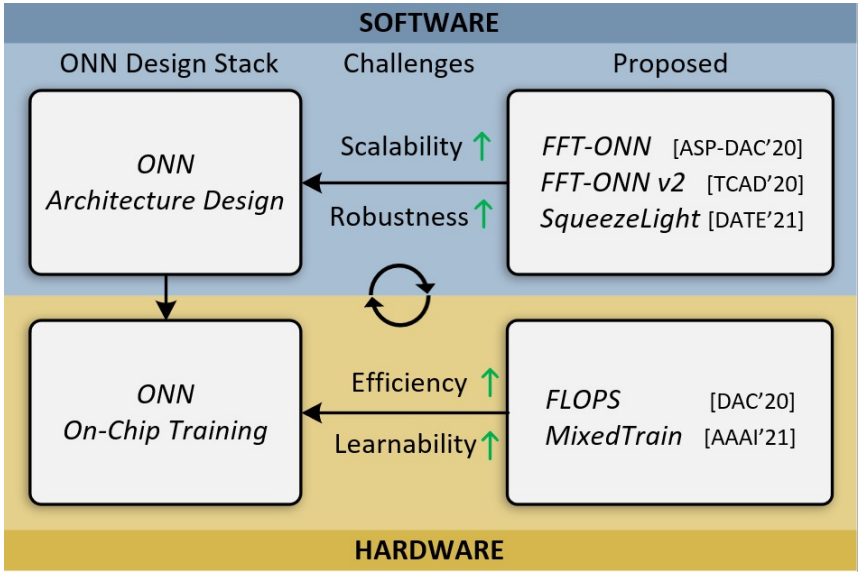
Experimental Results (Scalability & Power)

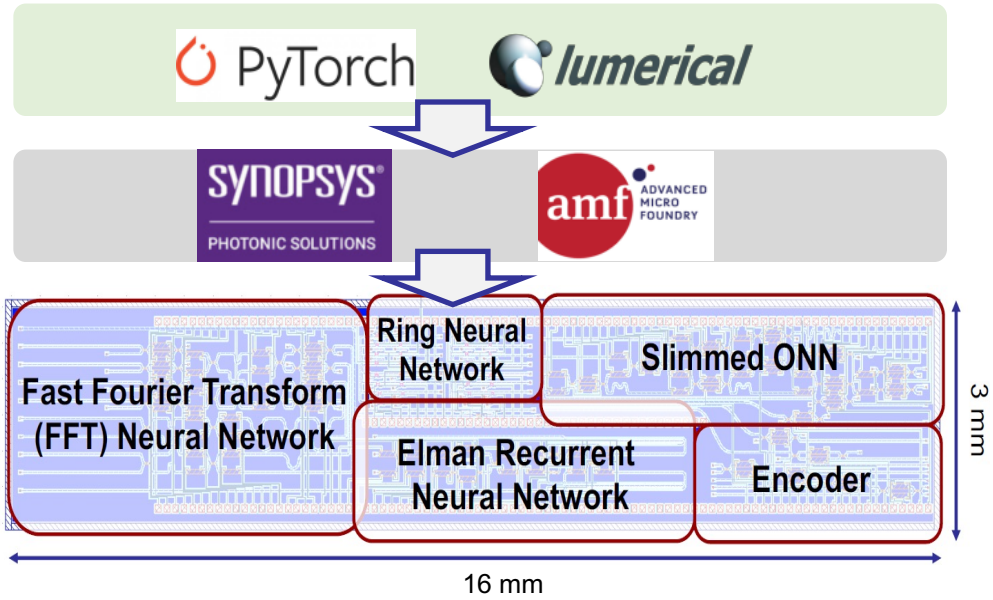
- Comparison with prior SOTA ZO optimizers on CNNs
  - Ours is the only one that stably converges
  - Highest average accuracy with ~2.5x better scalability (2500 MZIs vs. 1000 MZIs)
  - >95% lower power than naïve training ($\alpha = 1$) on CNN MNIST/CIFAR-10
  - >96% lower power than SOTA FLOPS [Gu+, *DAC*'20]



To Recap: Light in AI

- How to build ultra-fast (light-speed) and ultra-energy efficient optical neural accelerators with photonic integrated circuits
  - Software and hardware co-design is the KEY





Thanks!

Q & A?