DESIGN, AUTOMATION & TEST IN EUROPE
DATE 21
01 – 05 February 2021 · virtual conference
The European Event for Electronic System Design & Test

TEXAS
The University of Texas at Austin

SqueezeLight: Towards Scalable Optical Neural Networks with Multi-Operand Ring Resonators

Jiaqi Gu[1], Chenghao Feng[1], Zheng Zhao[2], Zhoufeng Ying[3], Mingjie Liu[1]
Ray T. Chen[1], David Z. Pan[1]

[1]ECE Department, University of Texas at Austin
[2]Synopsys, Inc., [3]Alpine Optoelectronics, Inc

This work is supported in part by AFOSR MURI

jqgu@utexas.edu; https://jeremiemelo.github.io

AI Acceleration: Challenges

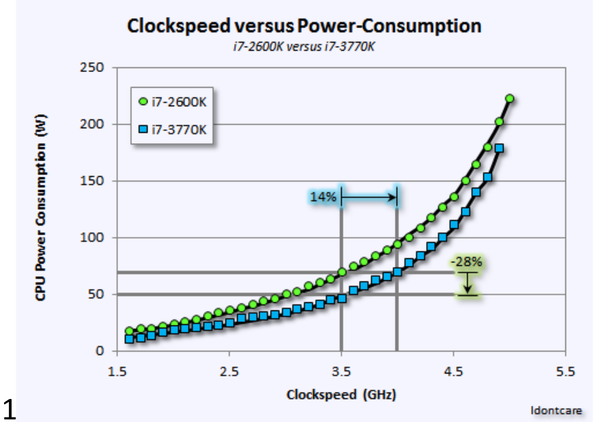
- ML models/dataset keep increasing -> more computations
  - Low latency
  - Low power
  - High bandwidth

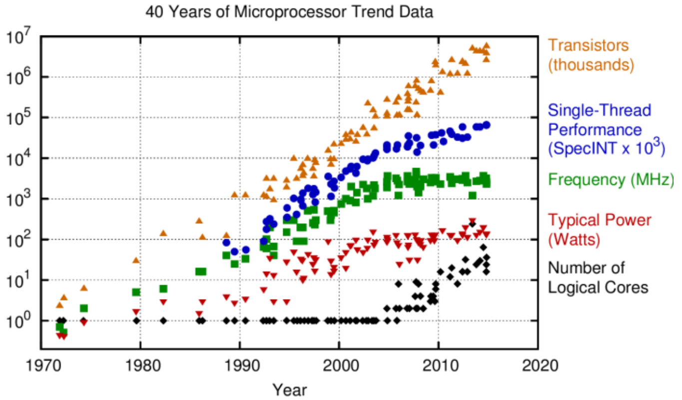
Autonomous Vehicle

Data Center

Edge Device

- Moore’s law is approaching its physical limits





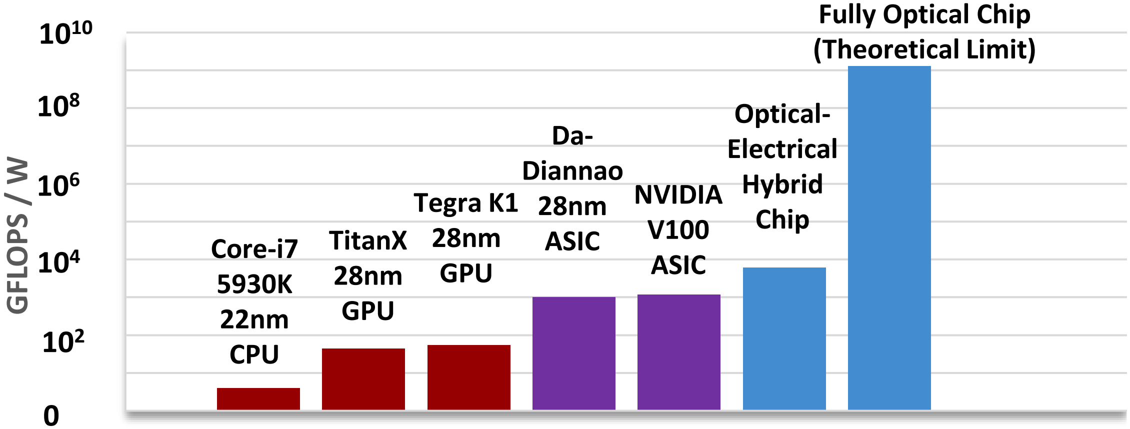
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

AI Acceleration: Opportunities

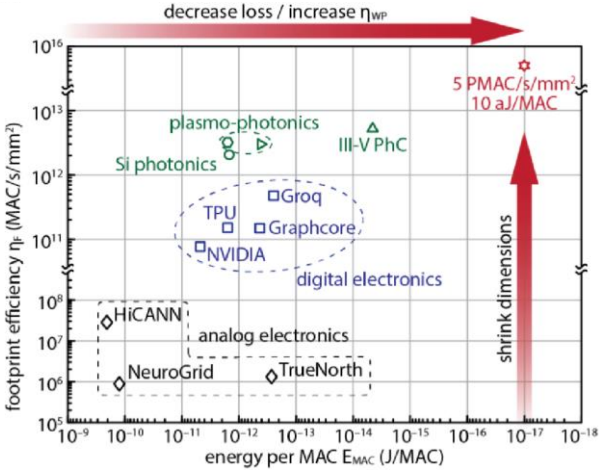
- Using light to continue Moore's Law
- Promising technology for next-generation AI accelerator

LIGHTELLIGENCE

LIGHTMATTER



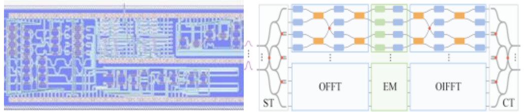
[Shen+, *Nature Photonics* 2017]



[Totovic+, *JSTQE* 2020]

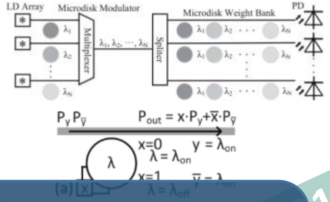
Optical Neural Networks (ONN)

- Emergence of photonic NNs
  - Ultra-low ps-level latency
  - Low energy consumption
- Compact design is bounded
  - 1 MAC per device

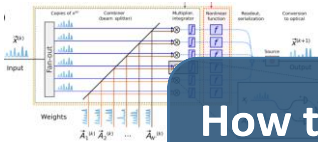


FFT-based optical neural network
[Gu+, ASPDAC2020]
[Gu+, TCAD2020]

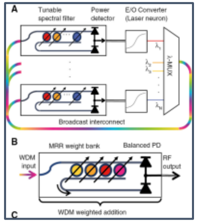
Holylight and Lightbulb: MRR&PCM
[Liu+, Zokaee+ DATE'2019, 2020]



Area Cost Still a Concern

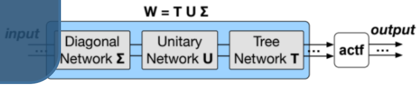


Quantum ONN
[Hamerly+, PhysRev2019]

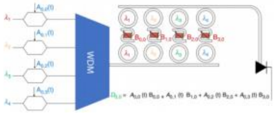


MRR Neural Network
[Brunner+, 2016]
[Tait+, SciRep 2017]

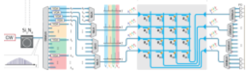
How to break through the area bound ?



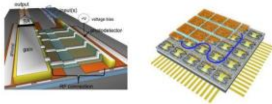
SlimmedONN
[Zhao+, ASPDAC2019]



PCM Xbar
[Miscuglio+, APR2020]

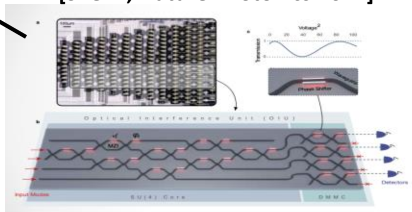


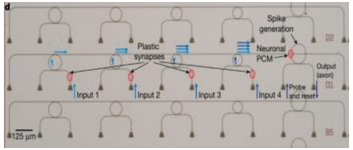
PCM Xbar
[Feldmann+, Nature2020]



Optical Spike NN
[Tait+, 2016]

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



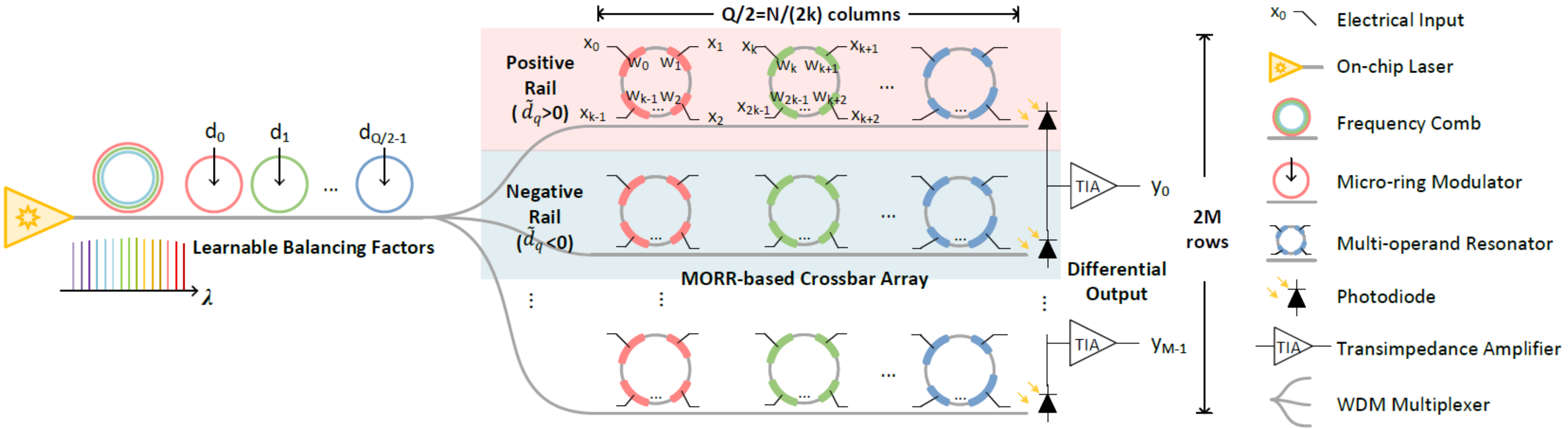


Spiking ONN: PCM
[Feldmann+, Nature 2019]

2016
2017
2018
2019
2020

Proposed SqueezeLight

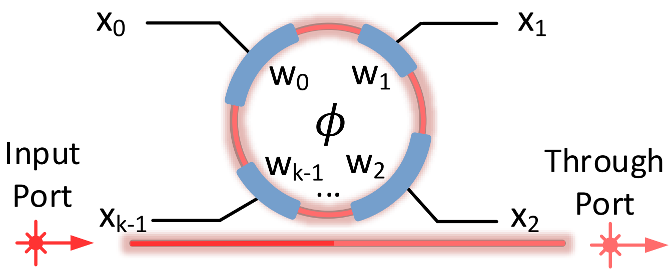
- `SqueezeLight`: ultra-compact MORR-ONN
  - **Scalability**: nonlinear neuron based on multi-operand ring resonators (MORR)
  - **Efficiency**: structured matrix with fined-grained structured pruning
  - **Robustness**: sensitivity-aware learning to overcome variations and crosstalk



Multi-Operand Ring Resonators

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

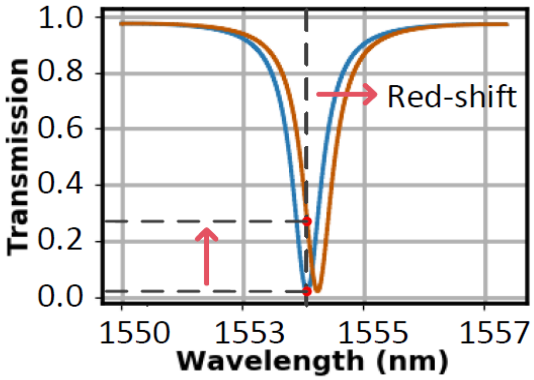
$$\text{Round-trip phase: } \phi \propto \sum_{i=0}^{k-1} w_i x_i^2$$



- Built-in nonlinearity
  - Half-Tanh-like nonlinear activation $f(\cdot) \in (0, 1)$
  - Tunable smoothness (r, a)

$$f(\phi) = \left| \frac{r - a\, e^{-j\phi}}{1 - ra\, e^{-j\phi}} \right|^2$$

$$OUT = f(\phi) \cdot in \propto f\left(\sum_{i=0}^{k-1} w_i x_i^2\right) \cdot IN$$

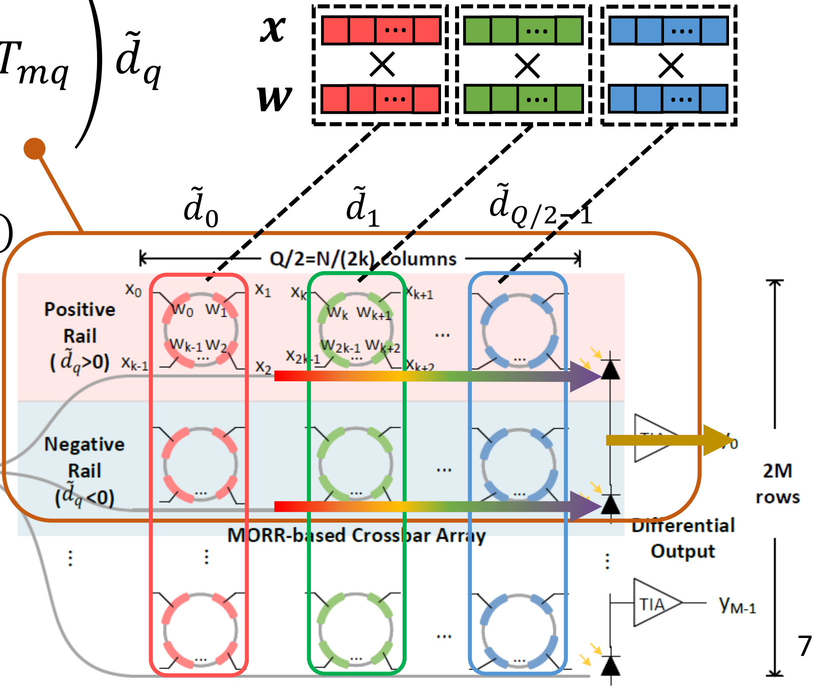


- No power consumption overhead
  - Same tuning range: half spectrum width, $\forall k$

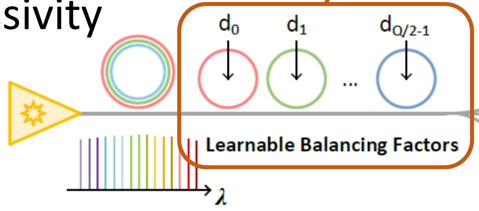
MORR-based ONN Architecture

- Nonlinear $M \times N$ MatMul in MORR crossbar array
- Differential rails support positive/negative neurons

$$y_m = \left(\sum_{q=0}^{Q/2-1} OUT_{mq} - \sum_{q=Q/2-1}^{Q-1} OUT_{mq} \right) \tilde{d}_q$$

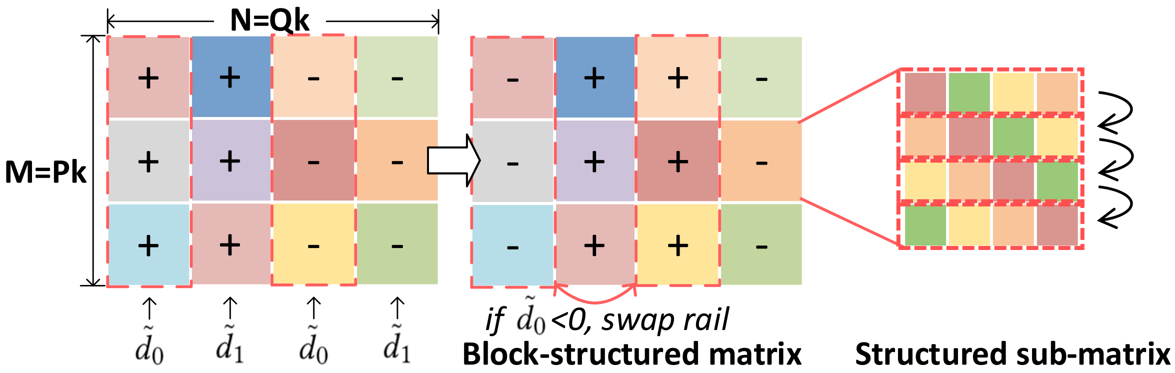


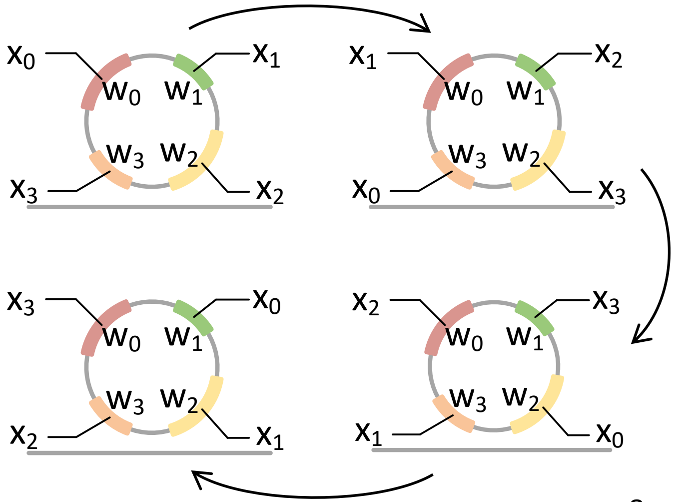
- Learnable balancing factors $(\tilde{d}_0, \tilde{d}_1, \cdots, \tilde{d}_{\frac{Q}{2}-1})$
  - Adaptive MORR output range
  - Enhanced expressivity



Area Reduction: Block-Squeezing

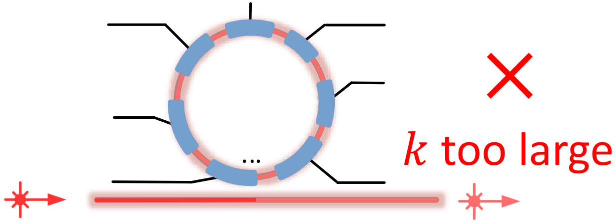
- Nonlinear $M \times N$ *Block-structured MatMul* in MORR crossbar array
- Squeeze a structured matrix into one MORR
  - Share weights in multiple rows → share the same MORR
  - Saves $k^2 \times$ device usage via input rotation
  - $k \times$ less weight storage
  - $2k \times$ fewer wavelengths



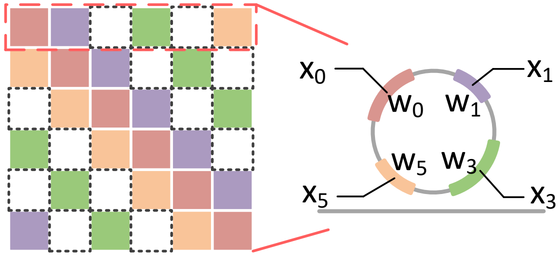


Sparsity Exploration: Fined-Grained Pruning

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



- Sparsify blocks via fine-grained structured pruning
  - 4-operand MORR $\longleftrightarrow$ 6×6 pruned block (33% sparsity)
  - 4-operand MORR $\longleftrightarrow$ 8×8 pruned block (50% sparsity)
  - Support larger blocks with small MORR
  - Pruning-aware training



Sparse structured sub-matrix

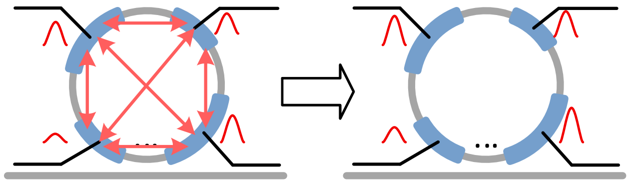
Robustness Boost: Sensitivity Optimization

- Non-ideal effects of MORRs
  - Individual phase drift

$$\Delta\phi \in \mathcal{N}(0, \sigma^2)$$

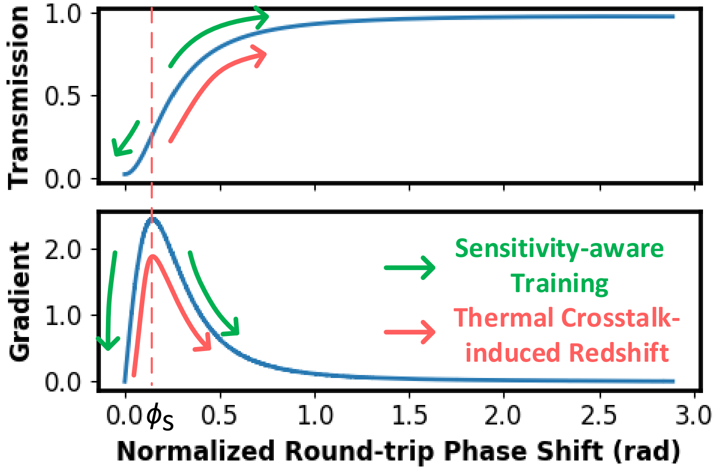
  - Intra-MORR crosstalk

$$\begin{pmatrix} \gamma_{0,0} & \gamma_{0,1} & \cdots & \gamma_{0,k-1} \\ \gamma_{1,0} & \gamma_{1,1} & \cdots & \gamma_{1,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k-1,0} & \gamma_{k-1,1} & \cdots & \gamma_{k-1,k-1} \end{pmatrix}$$



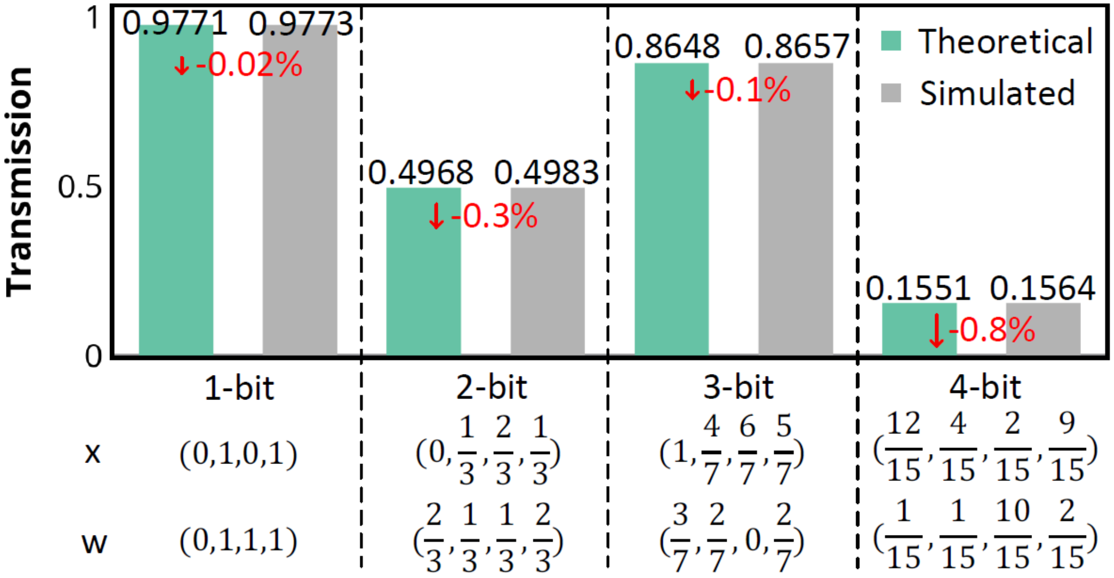
- Transmission sensitivity-aware regularization
  - Encourage phases with lower gradients

$$\mathcal{L} = \mathcal{L}_0(x; \boldsymbol{W}, \tilde{\boldsymbol{D}}, \boldsymbol{\Gamma}, \Delta\phi) + \alpha \sum_{l,m,q=0}^{L-1,M-1,Q-1} \nabla_\phi f(\hat{\phi}_{lmq} + \Delta\phi)$$



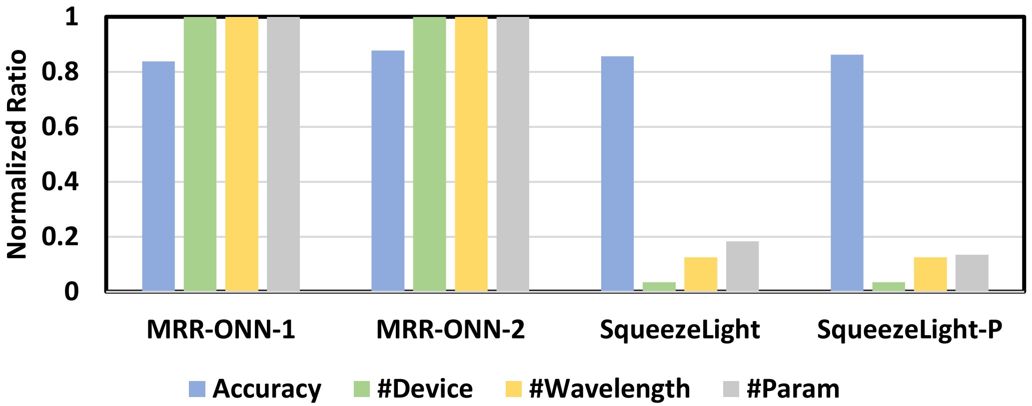
Fidelity Validation: Optical Simulation

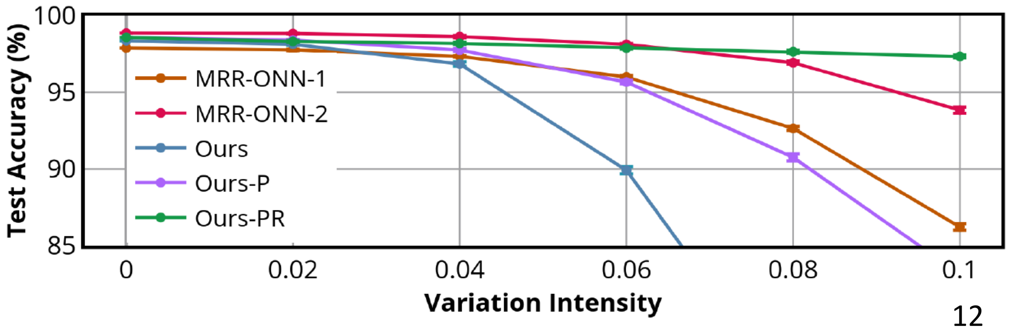
- 1- to 4-bit MORR neuron
- Optical simulation with Lumerical INTERCONNECT
- <1% relative modeling error



Comparison: Accuracy, Scalability, Robustness

- Compare with SoTA MRR-ONNs on MNIST, FMNIST CIFAR-10
  - MRR-ONN-1 [Liu+, DATE'2019]
  - MRR-ONN-2 [Tait+, SciRep'2017]
- Comparable expressivity
- 23×-32× less device usage
- 8× fewer wavelength usage
- >5× fewer parameters
  - 50% sparsity
  - No accuracy drop
- Better noise-robustness
  - Maintain > 97%

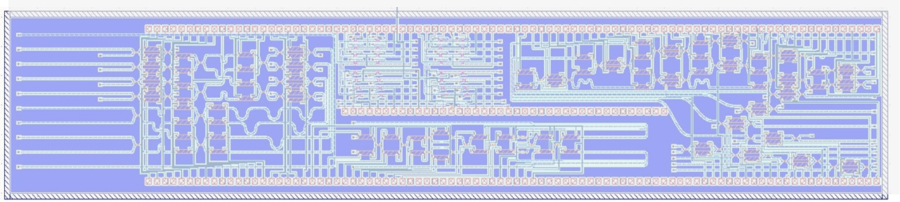




Conclusion and Future Work

- **New ONN Architecture**: Optical MORR-based neural architecture
- **Ultra-compact footprint**: 23×~32× fewer device usage, built-in nonlinearity
- **Better scalability**: 8× fewer wavelength usage
- **Better robustness**: ~4% higher accuracy under variations and crosstalk
- **Fewer parameters**: >5× fewer weight storage

- Future direction
  - Demonstrate more applications
  - Physical evaluation and testing on photonic neural chip tape-out



SYNOPSYS PHOTONIC SOLUTIONS

amf ADVANCED MICRO FOUNDRY

Thank You !

Q&A