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

TEXAS
The University of Texas at Austin

O^2NN: Optical Neural Networks with Differential Detection-Enabled Optical Operands

Jiaqi Gu[1], Zheng Zhao[2], Chenghao Feng[1], Zhoufeng Ying[3]
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://jeremoemelo.github.io

AI Acceleration: Challenges

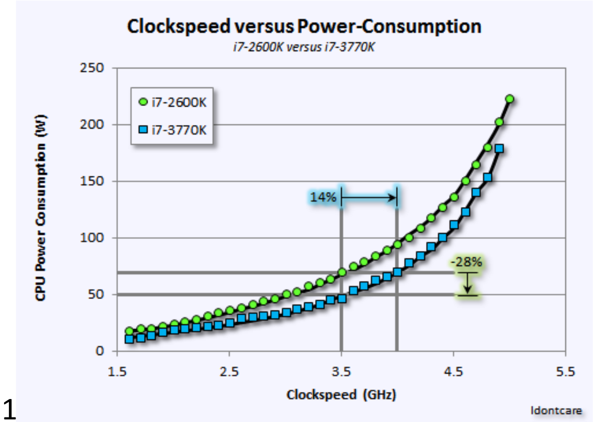
- ML models/dataset keep increasing -> more computations
  - Low latency
  - Low power
  - High bandwidth
  - Flexibility

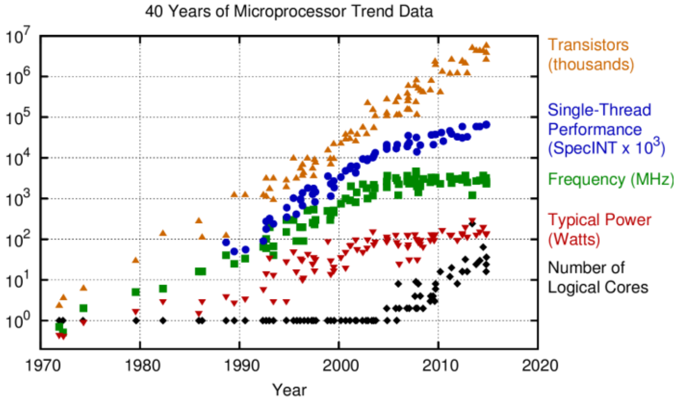
Autonomous Vehicle

Data Center

Edge Device

- Moore’s law is approaching its physical limits

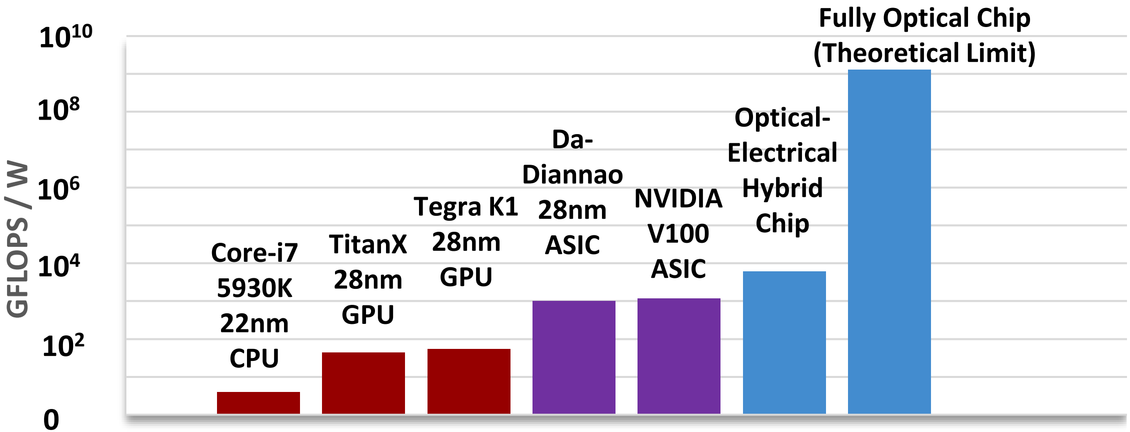




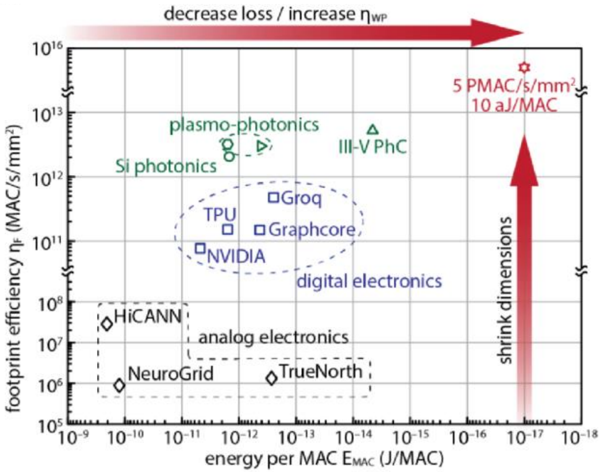
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

AI Acceleration: Opportunities

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



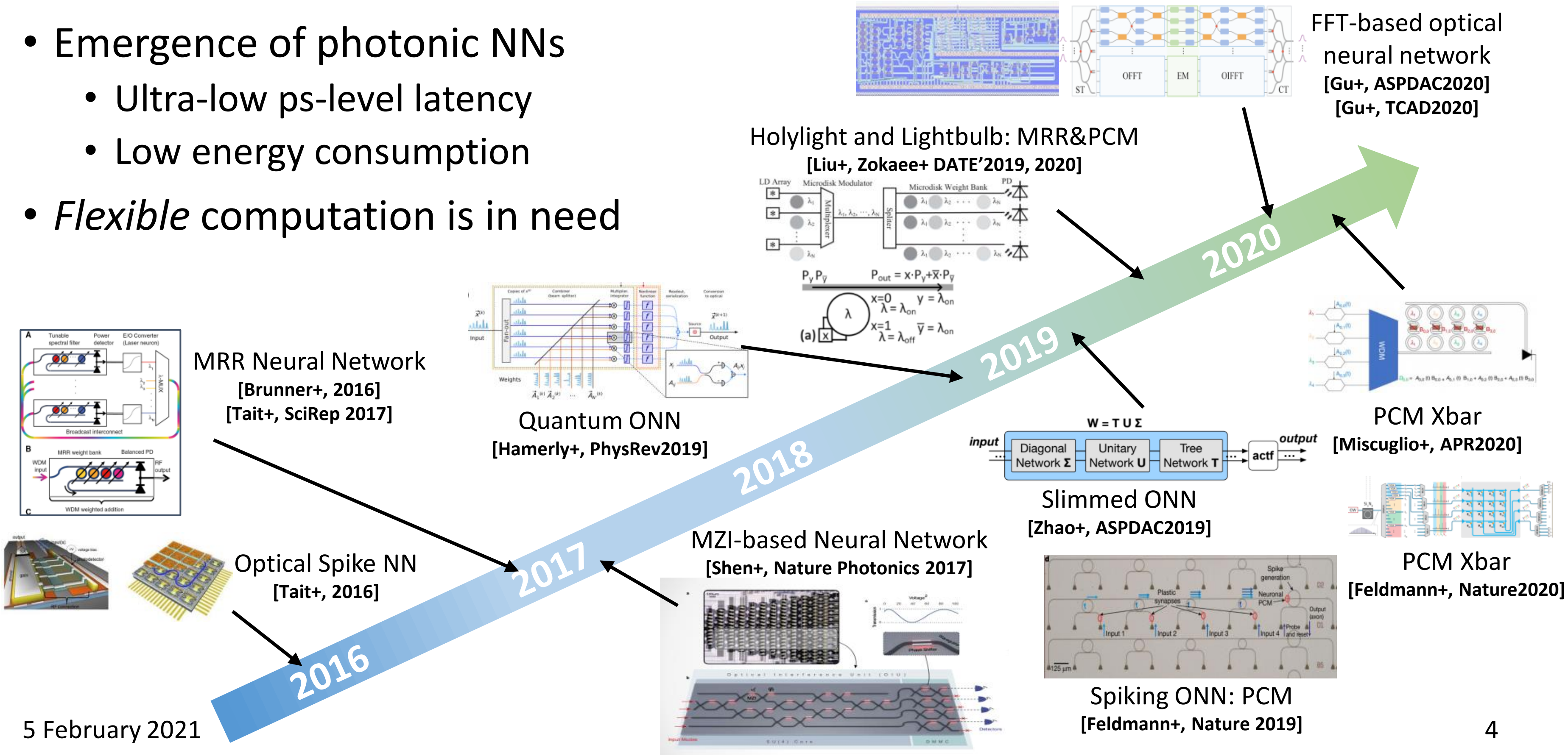
[Shen+, *Nature Photonics* 2017]



[Totovic+, *JSTQE* 2020]

Optical Neural Networks (ONN)

- Emergence of photonic NNs
  - Ultra-low ps-level latency
  - Low energy consumption
- *Flexible* computation is in need



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

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

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

Quantum ONN
[Hamerly+, PhysRev2019]

Optical Spike NN
[Tait+, 2016]

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

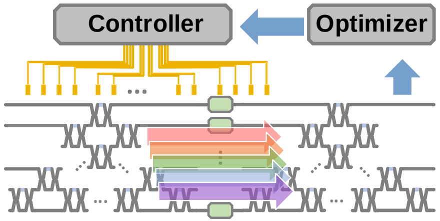
Slimmed ONN
[Zhao+, ASPDAC2019]

PCM Xbar
[Miscuglio+, APR2020]

PCM Xbar
[Feldmann+, Nature2020]

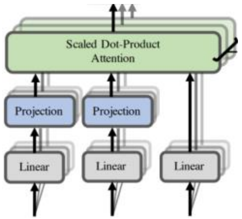
Spiking ONN: PCM
[Feldmann+, Nature 2019]

Flexibility Challenge



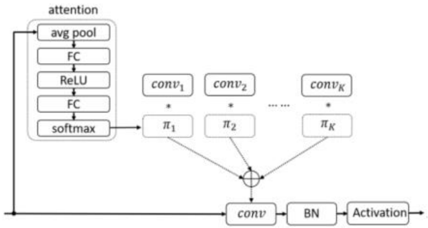
Training

[Gu+, *DAC*’20]



Attention

[Vaswani+, *NeurIPS*’17]



Dynamic Convolution

[Chen+, *CVPR*’20]

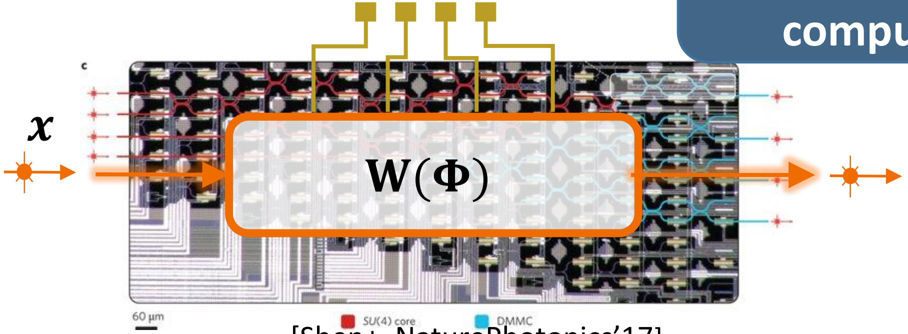
...

Stationary Design	Dynamic Design

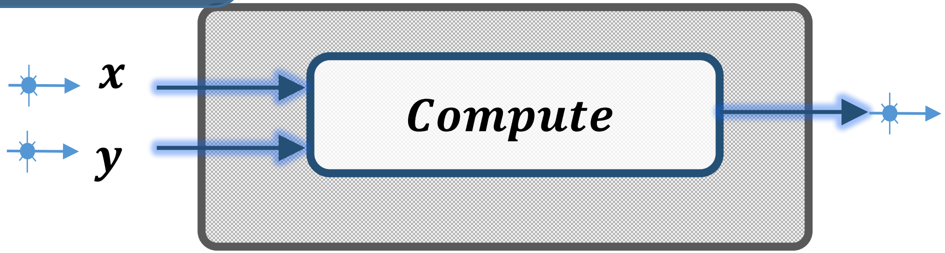
Implicit Nonlinear Encoding/Mapping

How to address *non-stationary* tensor computation ?

Explicit Linear Computing

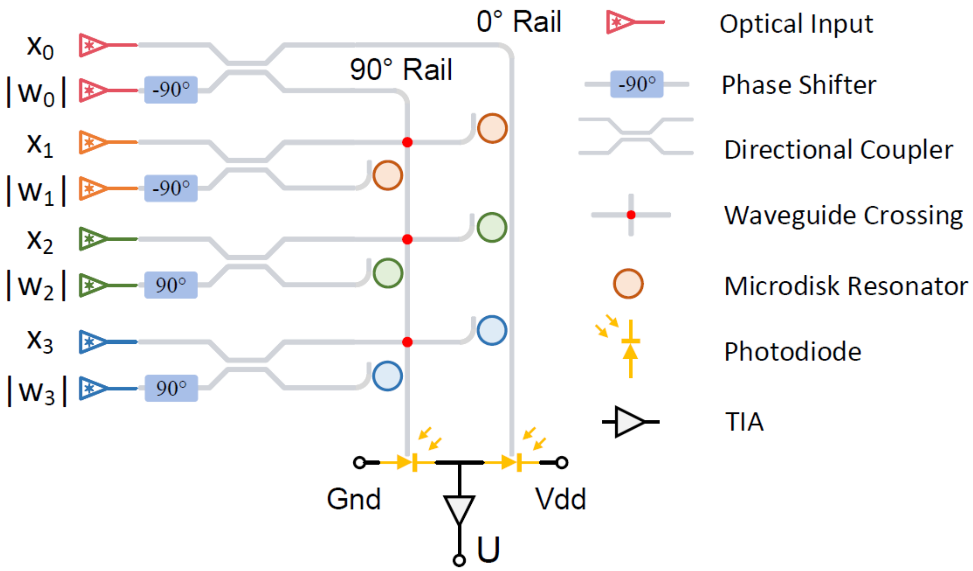


[Shen+, NaturePhotonics’17]



Proposed O^{2}NN

- O^{2}NN: Versatile ONN architecture with dynamic optical operands
  - **Flexibility**: differential detection-enabled fully-optical operands
  - **Expressivity**: extended optical weights and augmented quantization
  - **Robustness**: knowledge-distillation-based noise-aware training



Proposed Dot-Product Engine

- Interference between two optical signals

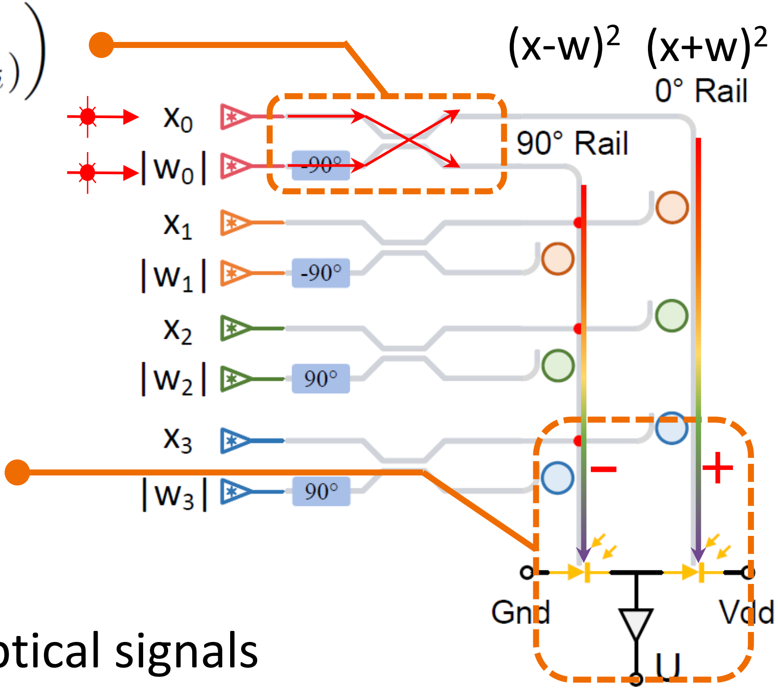
$$\begin{pmatrix} z_i^0 \\ z_i^1 \end{pmatrix} = \underbrace{\frac{1}{\sqrt{2}}\begin{pmatrix}1 & j\\ j & 1\end{pmatrix}}_{\text{directional coupler}} \underbrace{\begin{pmatrix}1 & 0\\ 0 & e^{-j\pi/2}\end{pmatrix}}_{\text{phase shifter}} \begin{pmatrix}x_i\\ w_i\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}x_i+w_i\\ j(x_i-w_i)\end{pmatrix}$$

- Dot-product via differential detection

$$\begin{pmatrix}I^0\\ I^1\end{pmatrix} = \frac{1}{2}\begin{pmatrix}\|\boldsymbol{x}+\boldsymbol{w}\|_2^2\\ \|j(\boldsymbol{x}-\boldsymbol{w})\|_2^2\end{pmatrix} = \frac{1}{2}\begin{pmatrix}\sum_{i=0}^{N-1}(x_i+w_i)^2\\ \sum_{i=0}^{N-1}(x_i-w_i)^2\end{pmatrix}$$

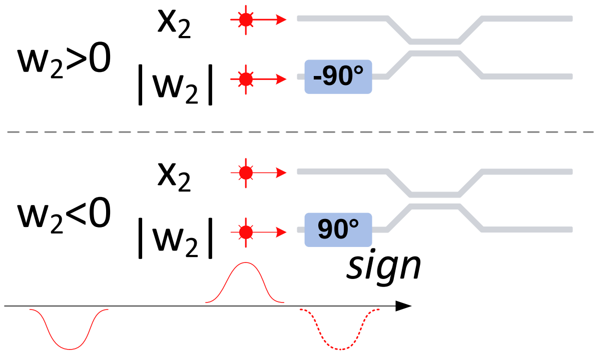
$$U = G(I_0 - I_1) = 2G\sum_{i=0}^{N-1} x_i w_i \propto \sum_{i=0}^{N-1} x_i w_i$$

- Both operands can be *high-speed dynamic* optical signals

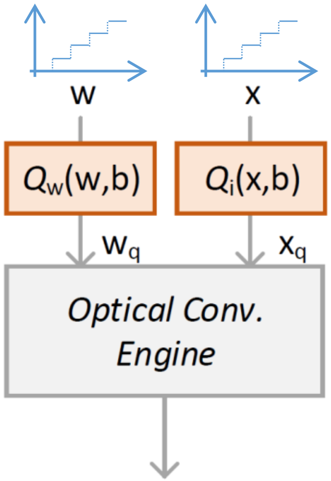


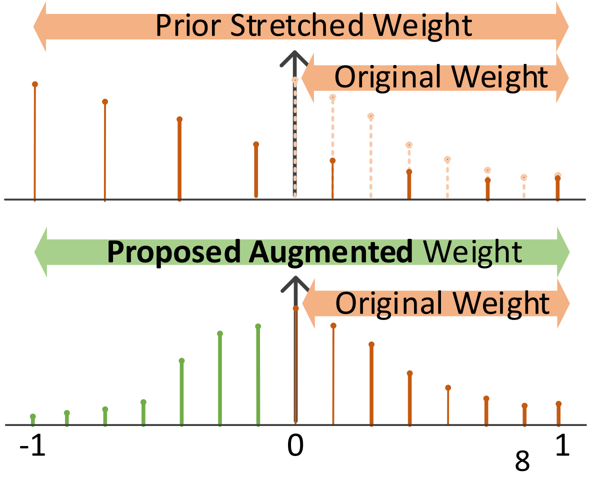
Expressivity Boost: Augmented Optical-Weight

- Optical-weight extension
  - Extra π phase shift on the input-port phase shifter
  - Can merge with the original -90° PS



- Augmented optical quantization
  - *b*-bit optical signal (non-negative)
  - (*b+1*)-bit equivalent weight (balanced)
  - Higher representability

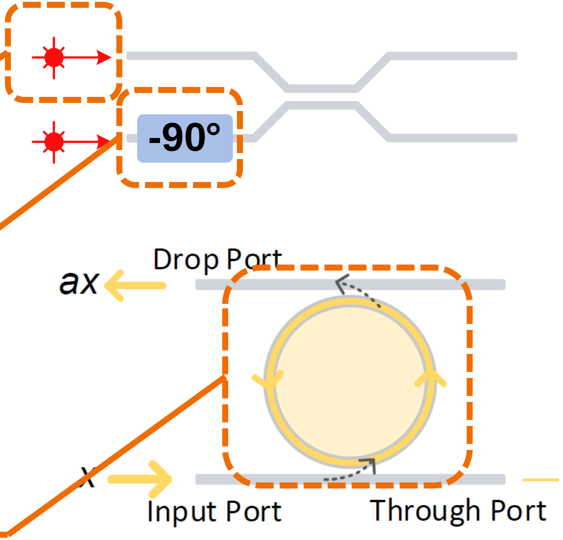


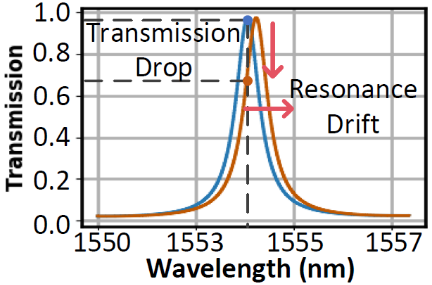


Robustness Analysis

- Dynamic input variations
  - Input signal-to-noise ratio: $SNR = \frac{\bar{P}(x)}{\bar{P}(\delta x)} = \frac{\mathbb{E}[x^2]}{\sigma^2}$
  - 10 dB for 40 Gb/s signal rate
  - $\hat{x}_i = (|x_i| + \delta x_i)e^{j(\frac{\pi}{2}+\delta\phi_i^d)}$

- Static device variations
  - Phase shifter drift: $\delta\phi_i^s \sim \mathcal{N}(0, \sigma_\phi^2)$
    - $\sin(\cdot)$ is stable at $\pm\frac{\pi}{2}$
  - MRR transmission drift: $\alpha \sim (1 - |\mathcal{N}(0, \sigma_\alpha^2)|)_+$
    - Spectrum is stable at $peak$
  - $\hat{U} \propto \sum_{i=0}^{N-1}\left(\frac{\alpha_i^0-\alpha_i^1}{4}(\hat{x}_i^2 + \hat{w}_i^2) - \frac{\alpha_i^0+\alpha_i^1}{2}\hat{x}_i\hat{w}_i \sin\phi_i\right)$





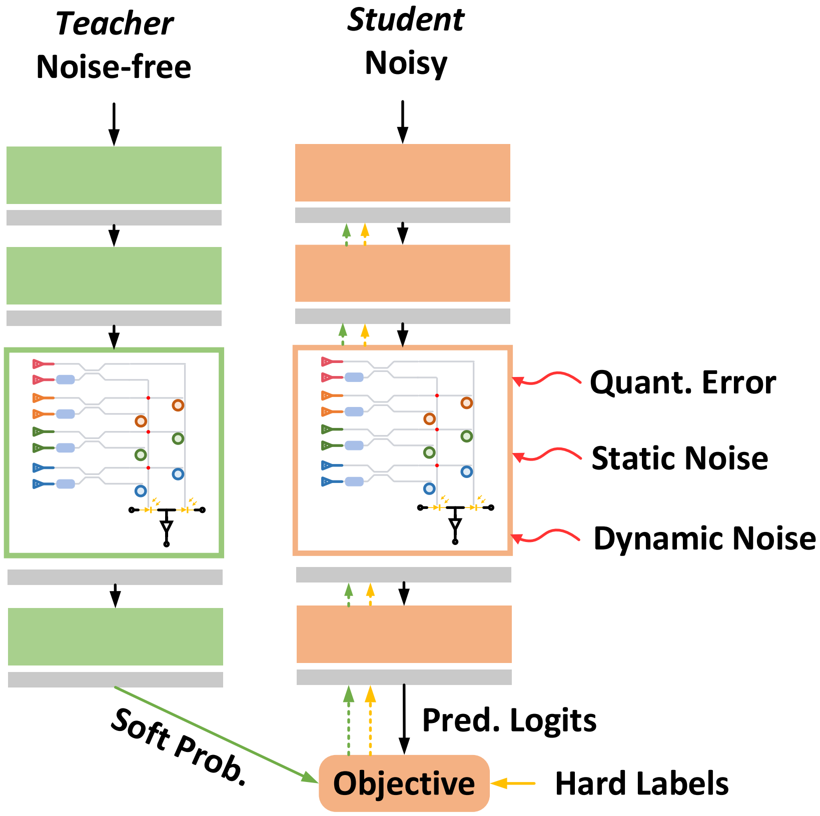
Robustness Solution: Knowledge Distillation

- Training ONNs with non-ideality modeling
- Pre-trained noise-free FP model as *teacher*
- *Student* model with noise and quantization
- Combined *KD* loss function
  - Cross-entropy with **hard** labels
  - KL-divergence with **soft** targets from *teacher*

$$\mathcal{L} = \beta T^2 \mathcal{D}_{KL}(q, p) + (1-\beta) H(y, \mathtt{softmax}(f_s))$$

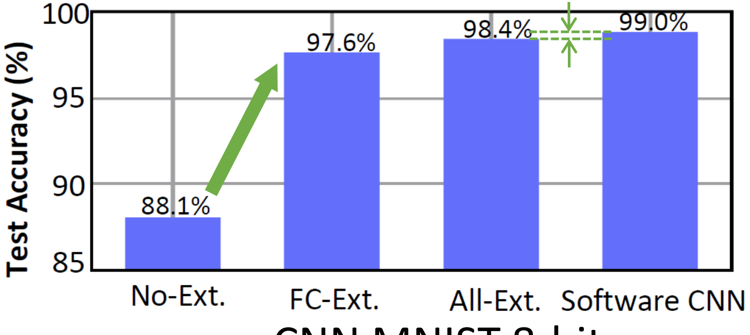
$$p = \frac{\exp(f_s/T)}{\sum \exp(f_s/T)}, \quad q = \frac{\exp(f_t/T)}{\sum \exp(f_t/T)},$$

- Better robustness to non-ideal effects



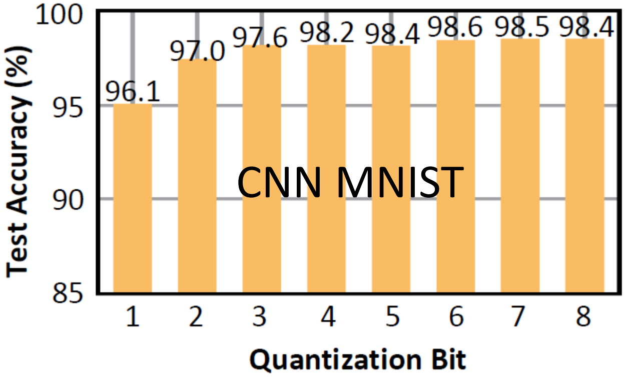
Experimental Results: Expressivity

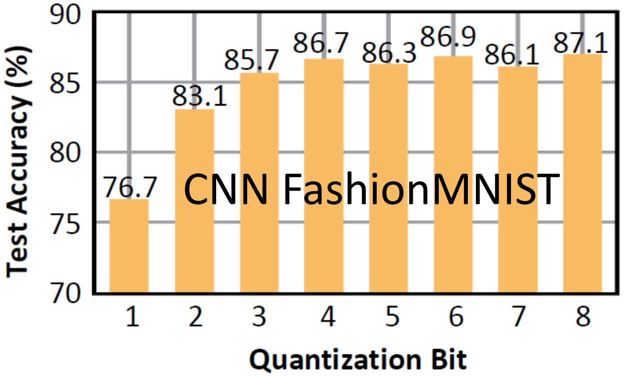
- Optical-weight extension
  - 10% better than model with positive weights
  - 0.6% accuracy drop compared with ideal model
  - Necessary for model expressivity



CNN MNIST 8-bit

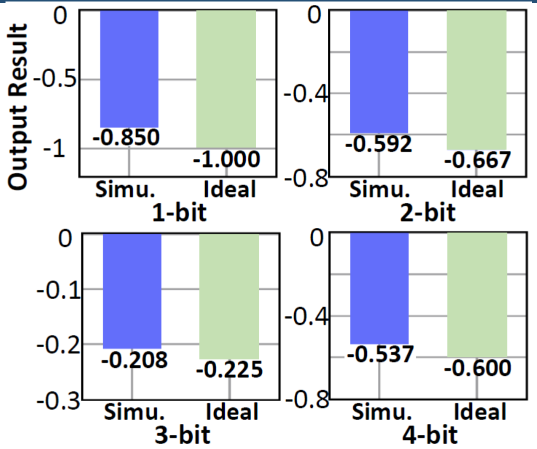
- Augmented optical quantization
  - ~1% accuracy drop with >3-bit



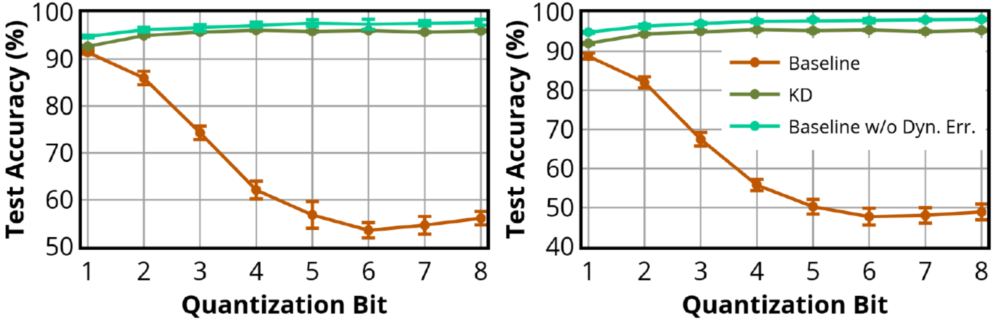


Experimental Results: Robustness

- Optical simulation
  - Lumerical INTERCONNECT with AMF PDKs
  - 10-15% dot-product error
- Knowledge-distillation based training
  - Only <3% accuracy drop under various bitwidths
  - 10%~20% more robust than prior MRR-ONN



O²NN



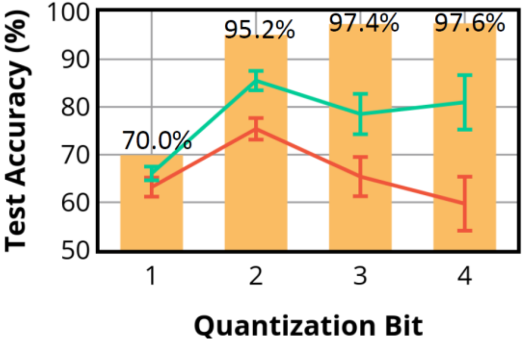
SNR=39.8dB

$\sigma_\phi = 0.04, \sigma_\alpha = 0.04$

SNR=31.6dB

$\sigma_\phi = 0.05, \sigma_\alpha = 0.05$

MRR-ONN



Conclusion and Future Work

- **New ONN engine with differential detection-enabled optical operands**
- **Flexibility**: Support dynamic, high-speed optical weights
- **Expressivity**: 2× more weight encoding with augmented quantization
- **Robustness**: 20%-30% more robust with knowledge-distillation

- Future direction
  - Integrate the fully-optical tensor core with dynamic NN architectures
  - Optimize the architecture with smaller device usage and footprint

Thank You !

Q&A