Towards Area-Efficient Optical Neural Networks: An FFT-Based Architecture

Jiaqi Gu, Zheng Zhao, Chenghao Feng, Mingjie Liu,
Ray T. Chen, David Z. Pan

ECE Department, The University of Texas at Austin

This work is supported in part by MURI

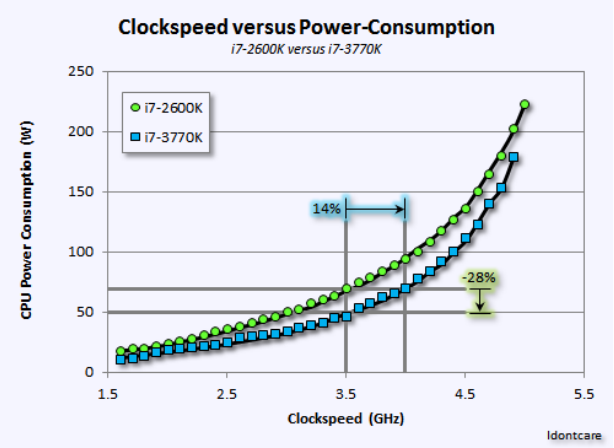
AI Acceleration and Challenges

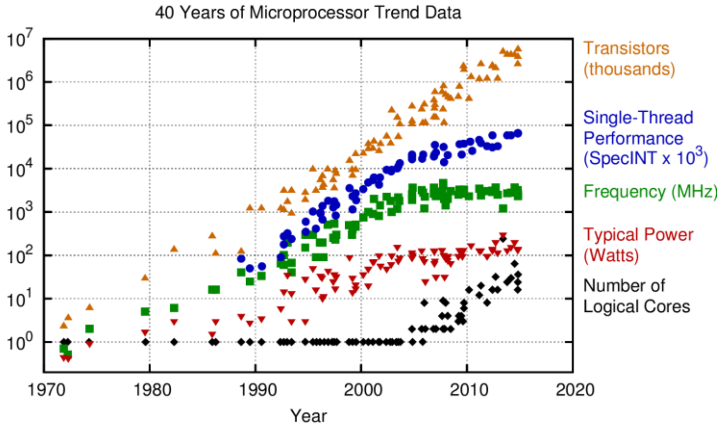
- ML models and dataset keep increasing -> more computation demands
  - Low latency
  - Low power
  - High bandwidth

Autonomous Vehicle

Data Center

- Moore’s law is challenging to provide higher-performance computations





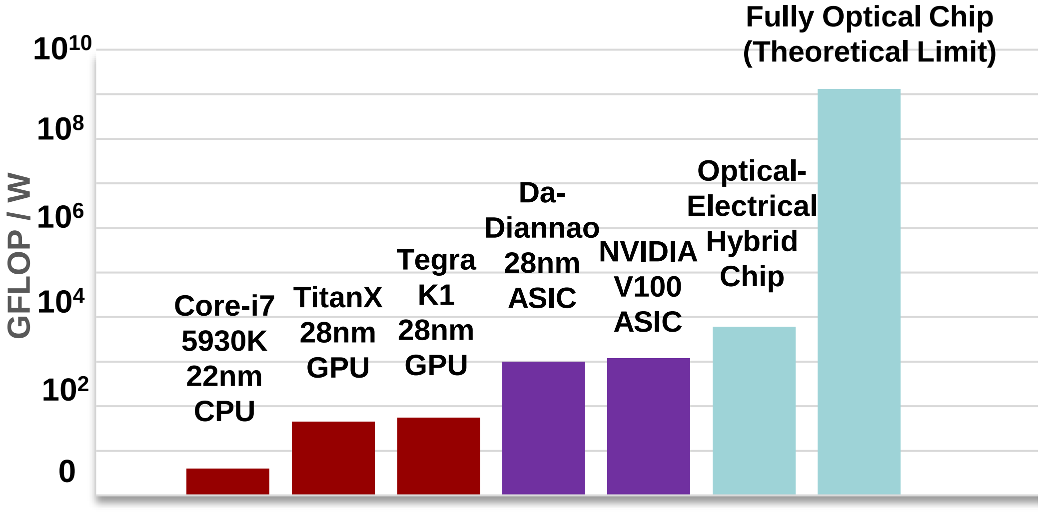
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

AI Acceleration and Challenges

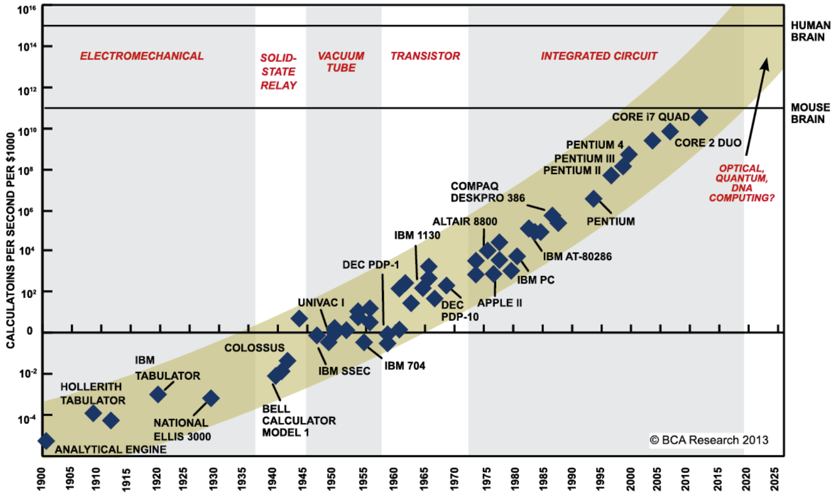
- Using light to continue Moore's Law
- Promising technology for next-generation AI accelerator

LIGHTMATTER

LIGHTELLIGENCE



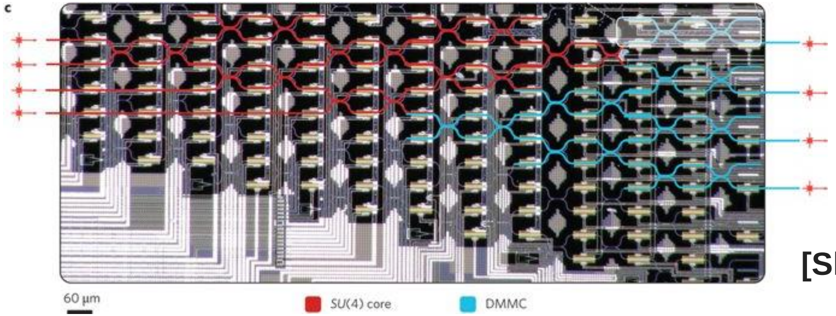
[Shen+, *Nature Photonics* 2017]

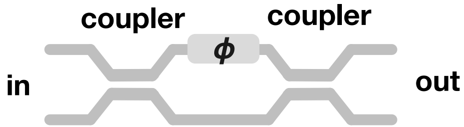


SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

Optical Neural Networks (ONN)

- Emergence of neuromorphic platforms for AI acceleration
- Optical neural networks (ONNs)
  - Ultra-fast execution speed (light in and light out)
  - >100 GHz photo-detection rate
  - Near-zero energy consumption if configured
- Unsatisfactory hardware area cost
  - Mach-Zehnder Interferometers (MZI) are relatively large
  - Previous architecture costs lots of MZIs (area-inefficient)
  - Previous architecture is not compatible with network pruning



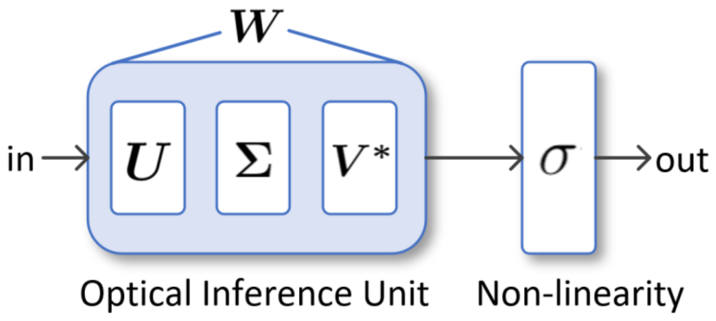


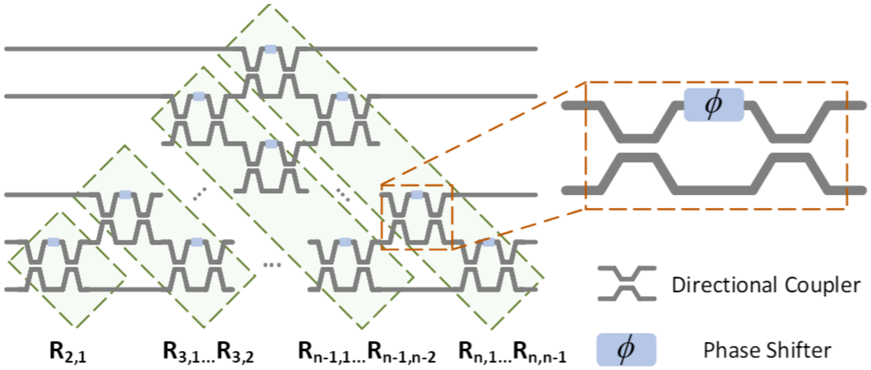
[Shen+, *Nature Photonics* 2017]

Previous MZI-based ONN Architecture

- Map weight matrix to MZI arrays
- Singular value decomposition
  - $W = U\Sigma V^*$
  - **U and V*** are square unitary matrices
  - **Σ** is diagonal matrix
- Unitary group parametrization
  - $$U(n) = D \prod_{i=n}^{2} \prod_{j=1}^{i-1} R_{ij}$$
  - R_{ij} is planar rotation matrix
  - R_{ij} with phase ϕ can be implemented by an MZI

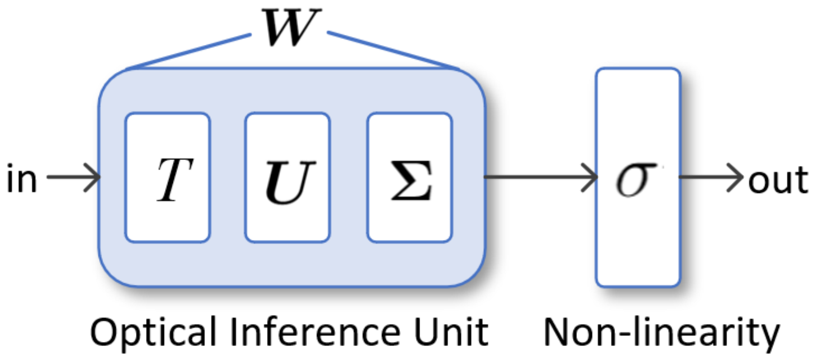
$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

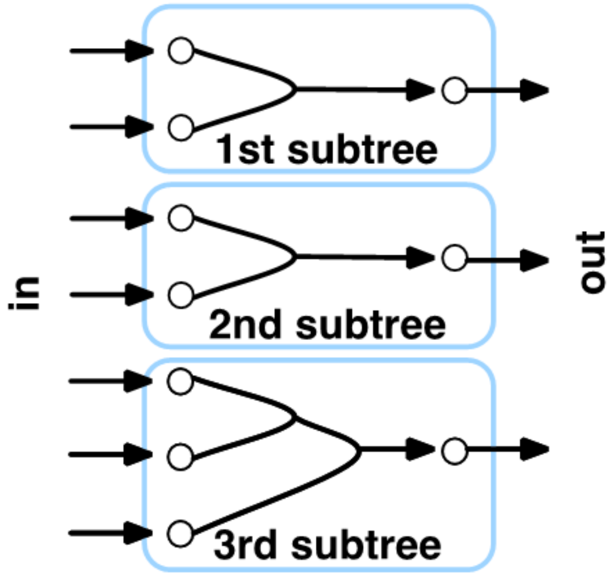




Previous MZI-based ONN Architecture

- Slimmed ONN architecture [ASPDAC'19 Zhao+]
- TUΣ decomposition
  - **T** is a sparse tree network for dimension matching
  - **U** is a square unitary matrix
  - **Σ** is diagonal matrix
- Use less # of MZIs
- Limits: only remove the smaller unitary

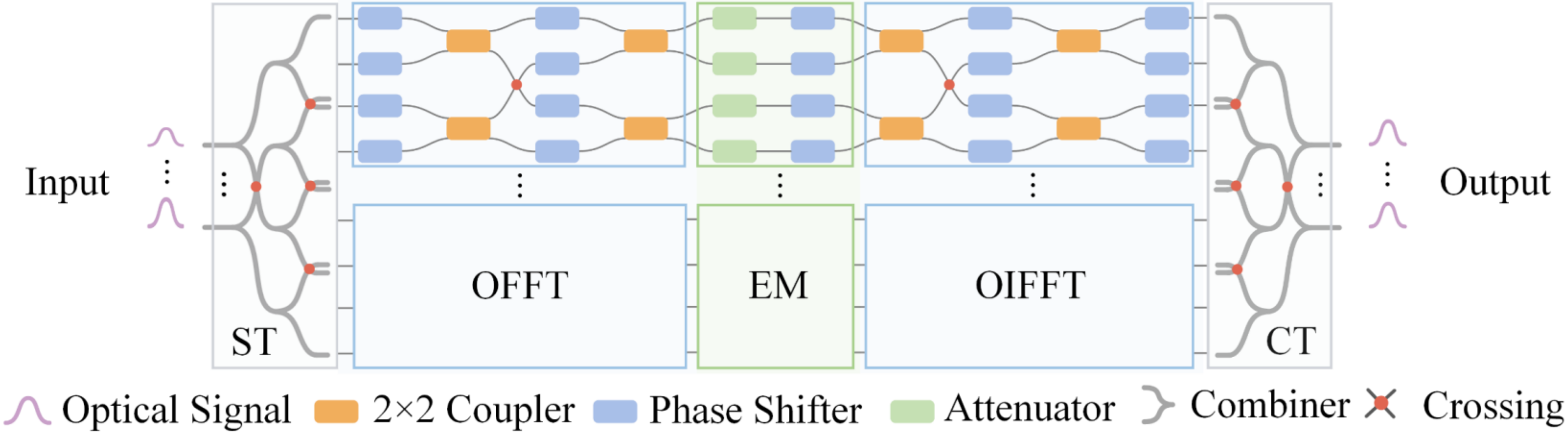




[ASPDAC'19 Zhao+]

Our Proposed FFT-ONN Architecture

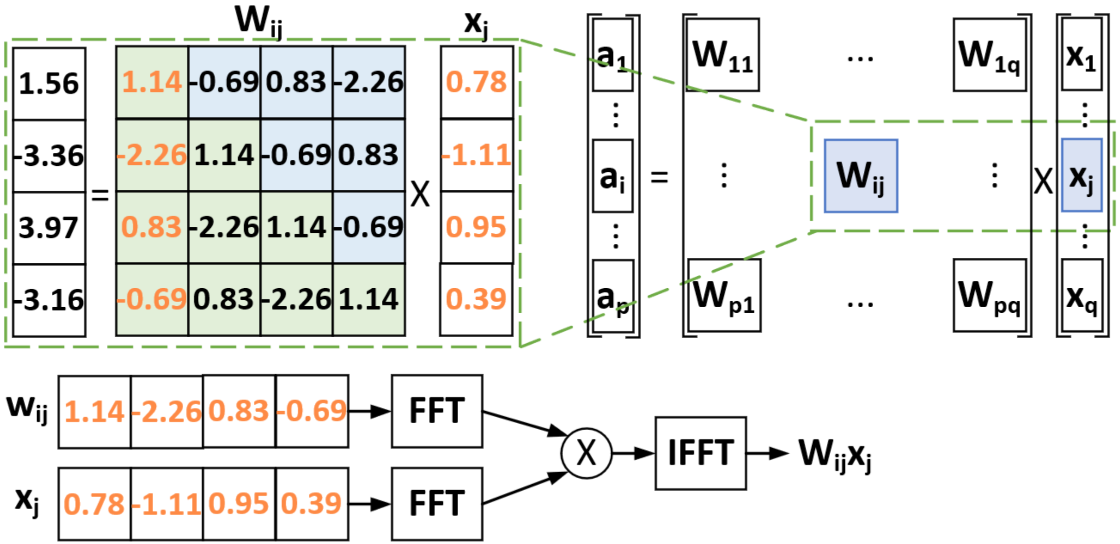
- Efficient **circulant matrix multiplication** in Fourier domain
- 2.2~3.7X area reduction
- Without accuracy loss



ST/CT: Splitter/Combiner tree (*Signal Fanout/Accumulation*)
OFFT/OIFFT: Optical FFT/IFFT (*Fourier Domain Transform*)
EM: Element-wise multiplication (*Weight Encoding in Fourier Domain*)

Block-circulant Matrix Multiplication

- Not general matrix multiplication
- Block-circulant matrix: each *k* x *k* block is a circulant matrix



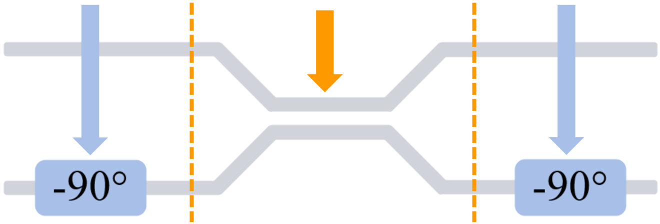
- Efficient algorithm in Fourier domain $y = Wx \Longleftrightarrow y = \mathcal{F}^{-1}(\mathcal{F}(w) \odot \mathcal{F}(x))$
- Comparable expressiveness to classical NNs. [ICLR'18 Li+]

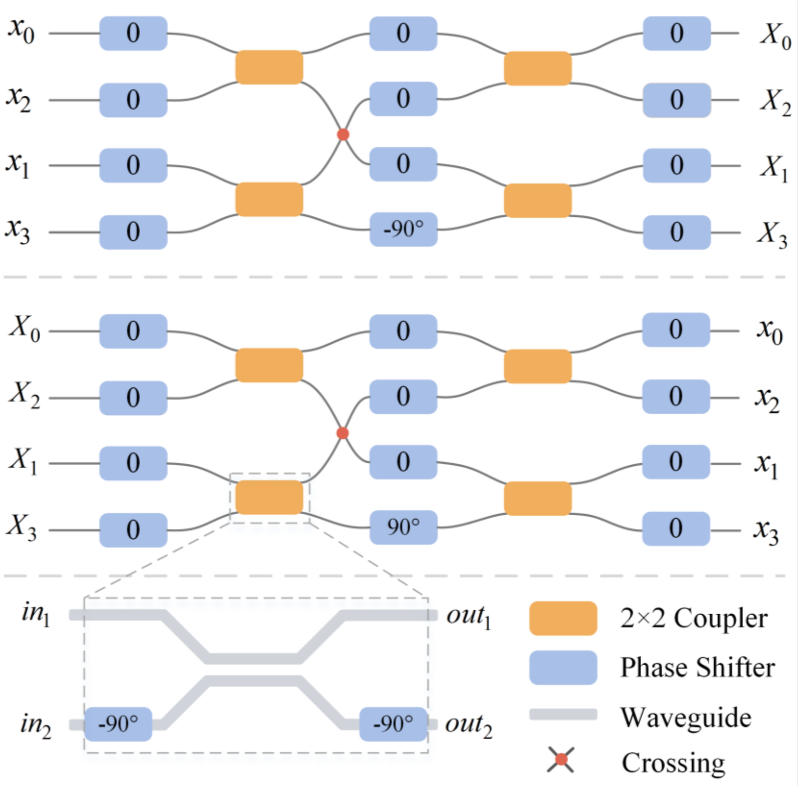
OFFT/OIFFT $\mathcal{F}(x)$

- Basic structure for 2-point FFT
  - 2×2 directional coupler
  - $-\pi/2$ phase shifter

$$\begin{pmatrix} out_1 \\ out_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} in_1 + in_2 \\ in_1 - in_2 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & -j \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -j \end{pmatrix} \begin{pmatrix} in_1 \\ in_2 \end{pmatrix}$$

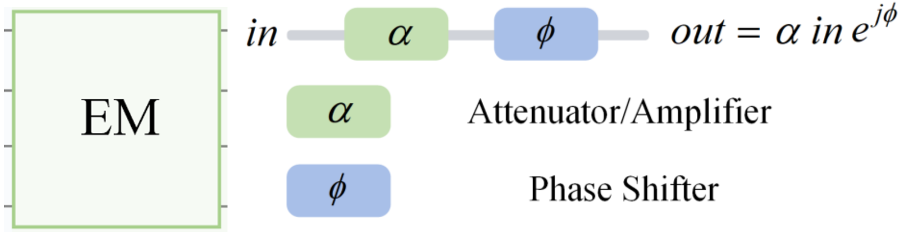


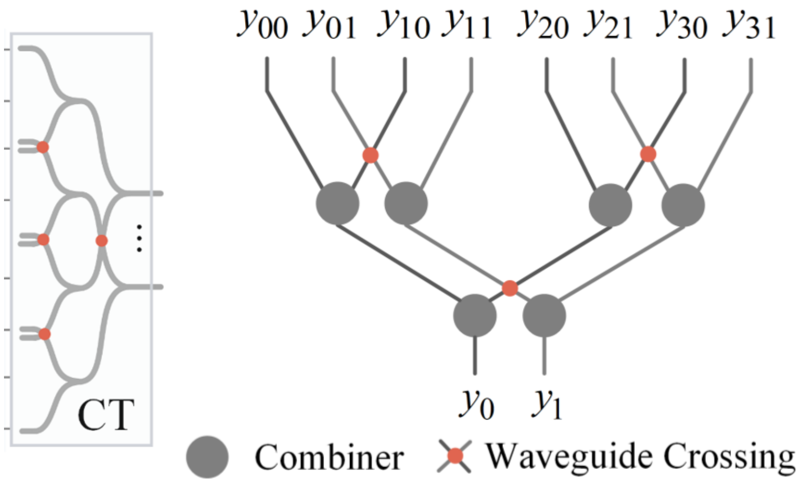


Weight Encoding $\mathcal{F}(w) \odot \mathcal{F}(x)$

- Multiplication in Fourier domain
  - Attenuator: magnitude modulation
  - Phase shifter: phase modulation
- Enable online/on-chip training
  - No complicated decomposition
  - Gradient backprop. friendly
- Splitter tree: fanout
- Combiner tree: accumulation
  - Fewer # of crossings: $O(n)$

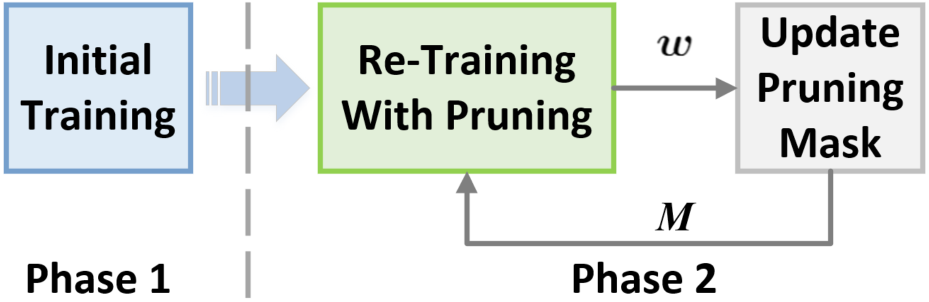
$$\alpha e^{\phi} \cdot I_i e^{\phi_i} = \alpha\, I_i e^{\phi_i + \phi}$$



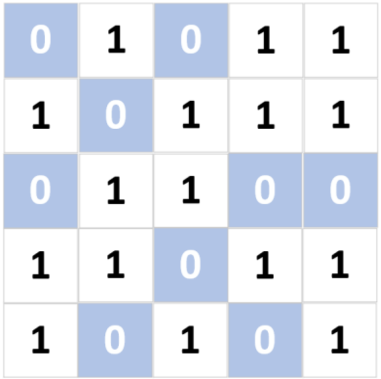


ONN Structured Pruning Flow

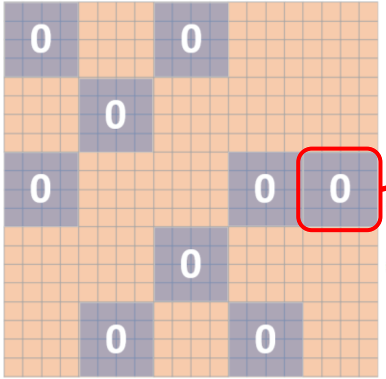
- Two-phase structured pruning
  - Group lasso regularization
  - Save **30% - 40%** components
  - Without accuracy loss (<0.5%)



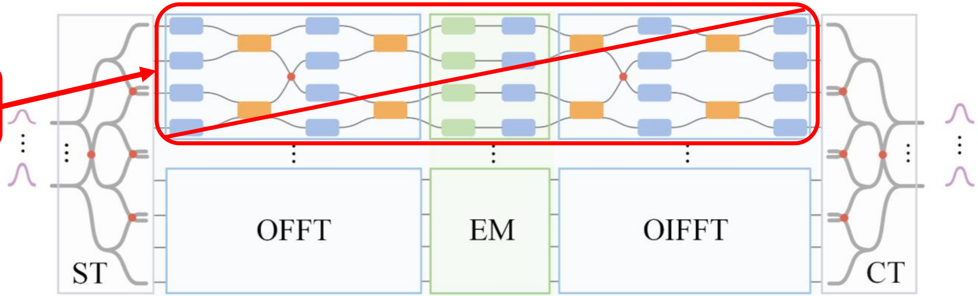
Masked 4 x 4 block eliminates the corresponding hardware



Pruning Mask M

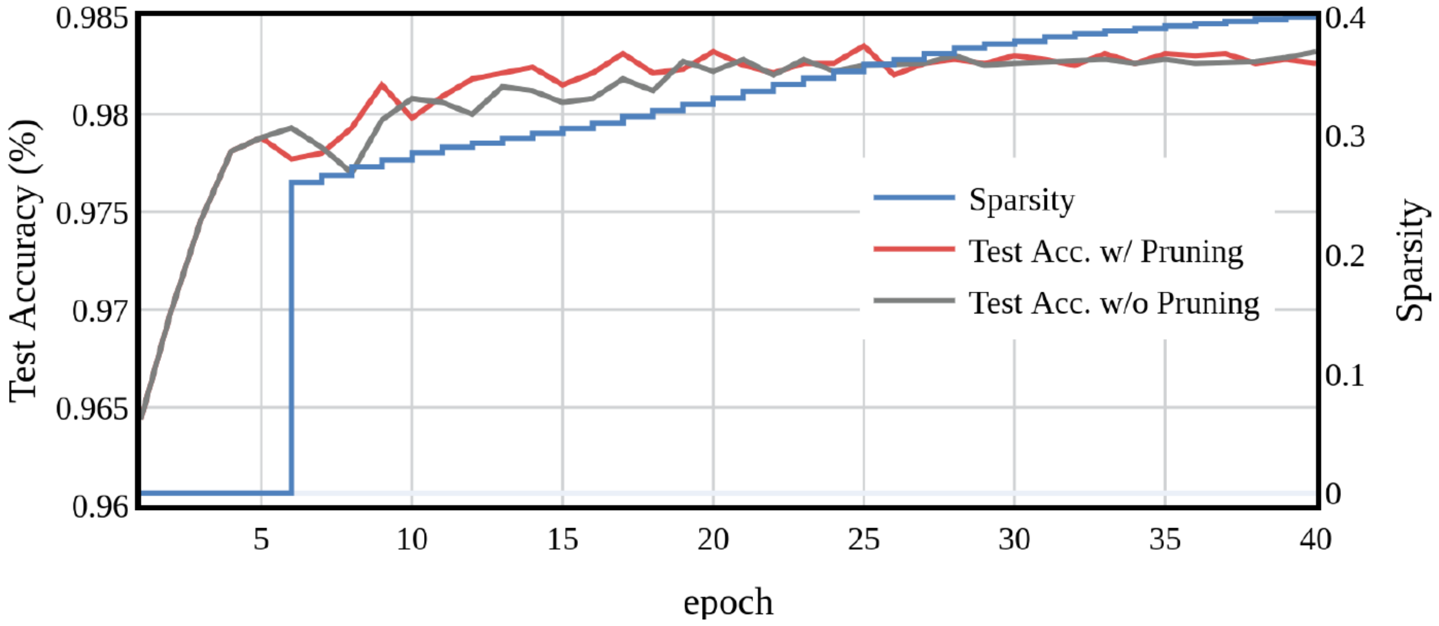


Masked Weight



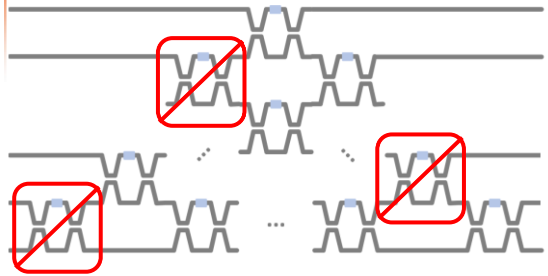
Training Curve

- Same convergence speed as *w/o pruning*
- Negligible accuracy loss (<0.5%)



Pruning-compatibility Comparison

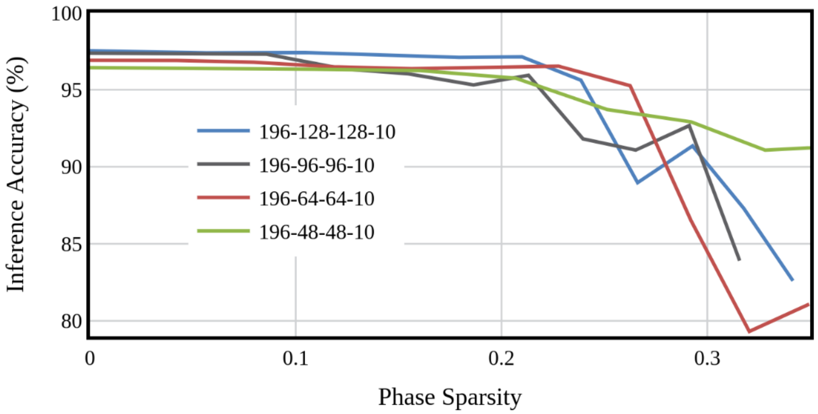
- Indirect and complicated



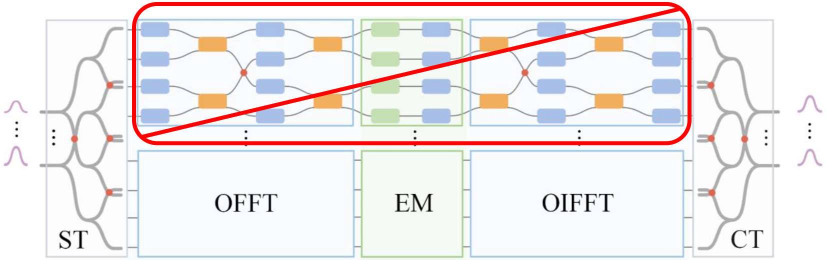
$$W = U\Sigma V^*$$

$$U(n) = D \prod_{i=n}^{2} \prod_{j=1}^{i-1} R_{ij}$$

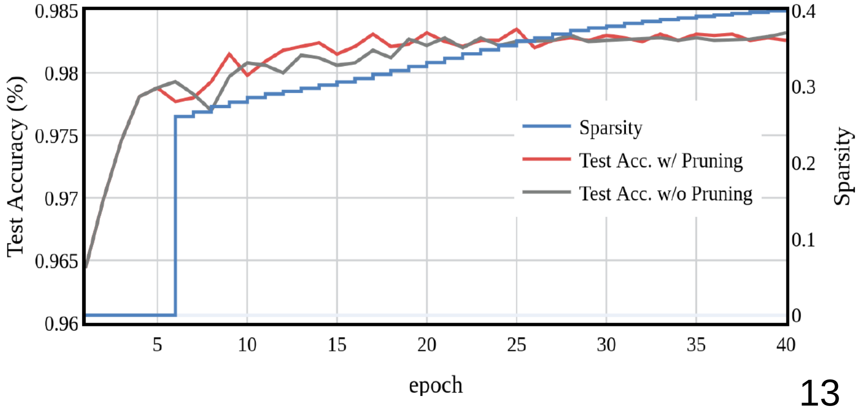
- Severe degradation



- Direct pruning



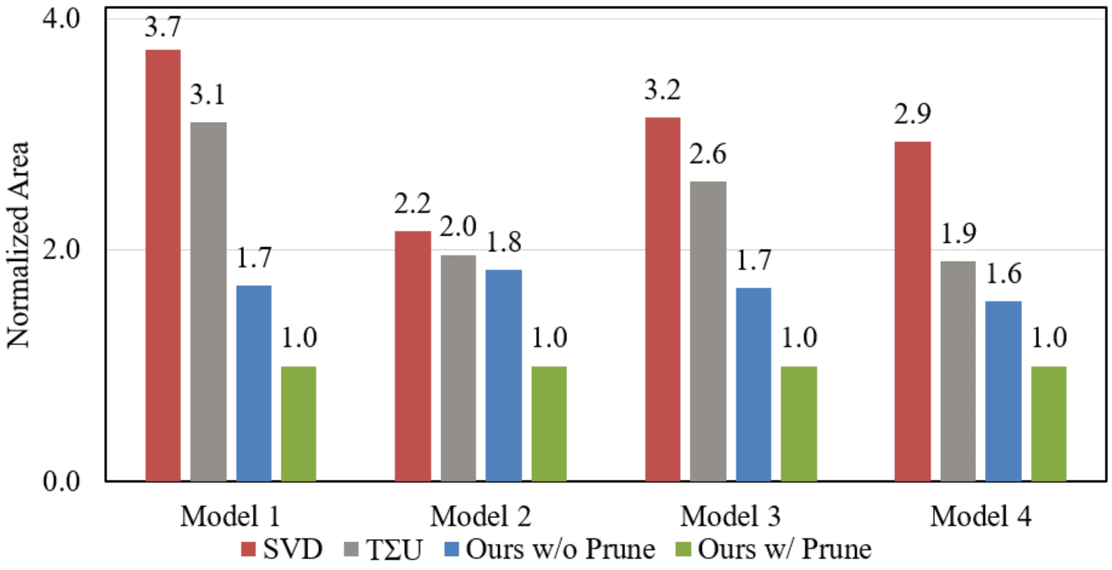
- No accuracy loss



Experimental Results

- 2.2~3.7X area cost reduction on various network configurations
- Similar accuracy (<0.5% diff)

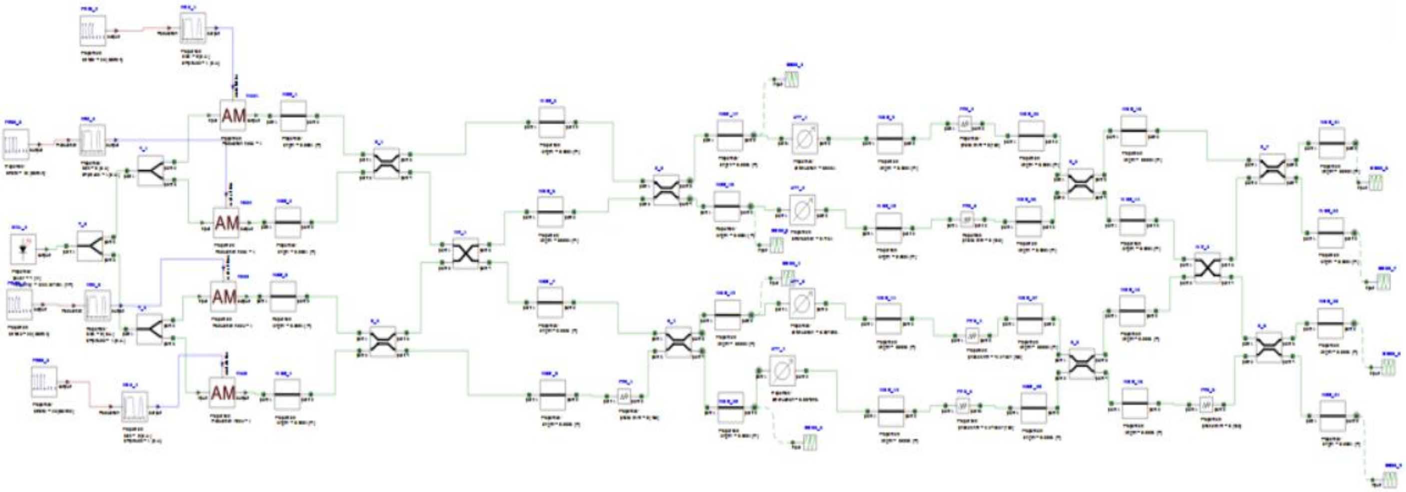
$$O(m^2 + n^2) \longrightarrow O(\frac{mn}{k}\log_2 k)$$



SVD: [Shen+, *Nature Photonics* 2017]   TΣU: [Zhao+, ASPDAC 2019]

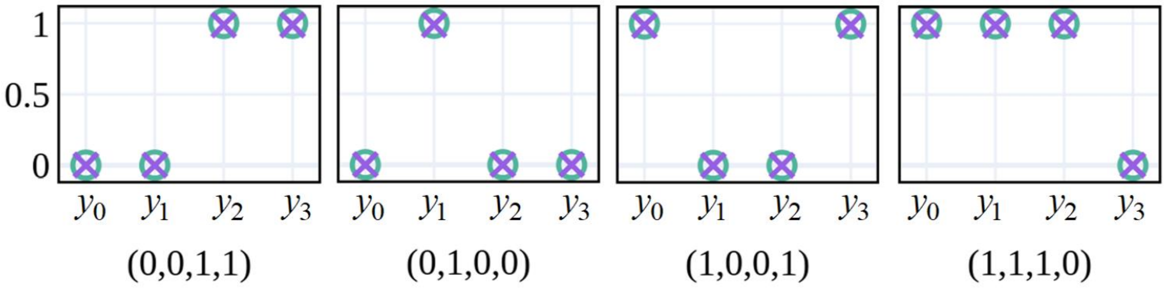
Simulation Validation

- Lumerical INTERCONNECT tool
- Device-level numerical simulation

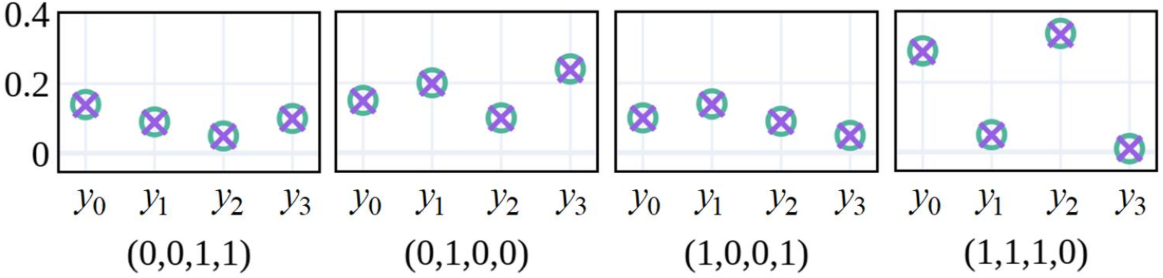


Simulation Validation

- Lumerical INTERCONNECT simulation (<1.2% maximum error)
  - 4 x 4 identity projection



  - 4 x 4 circulant matrix multiplication



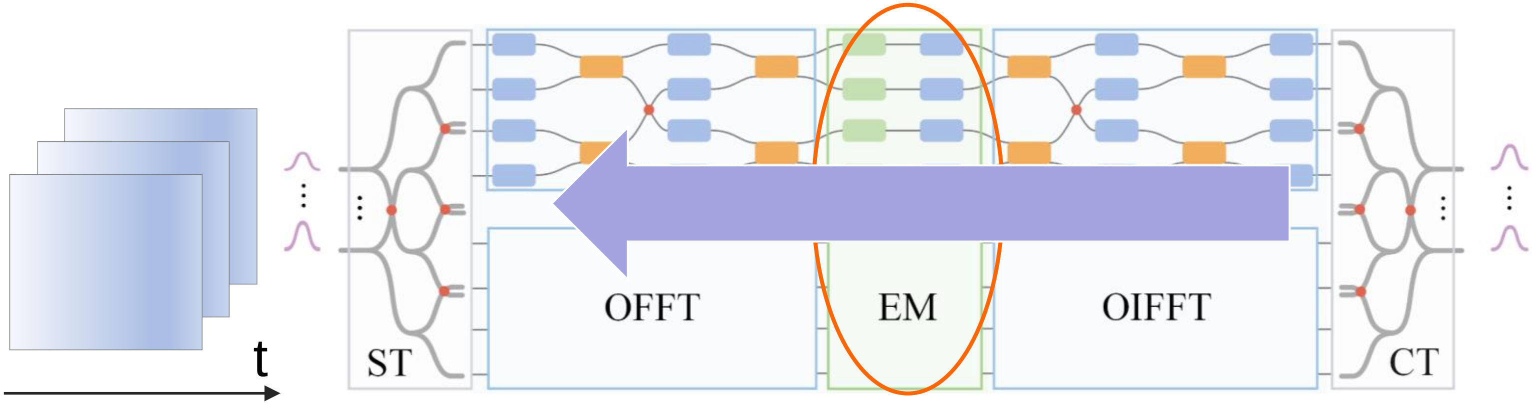
FFT-based ONN Summary

- A new ONN architecture
  - Without using MZI
  - **2.2X ~ 3.7X** lower area cost
  - Near-zero accuracy degradation

- Fourier-domain ONN
  - Efficient neuromorphic computation using Fourier optics
  - Better compatibility to NN compression
  - Enable on-chip learning

Extension and Potential

- Beyond classical real matrix multiplication
  - Enhanced expressiveness w/ latent weights in the complex domain
- Beyond 1-D multi-layer perceptron
  - Extensible to 2-D frequency-domain optical convolution neural network
- Beyond inference acceleration
  - Efficient on-chip training / self-learning

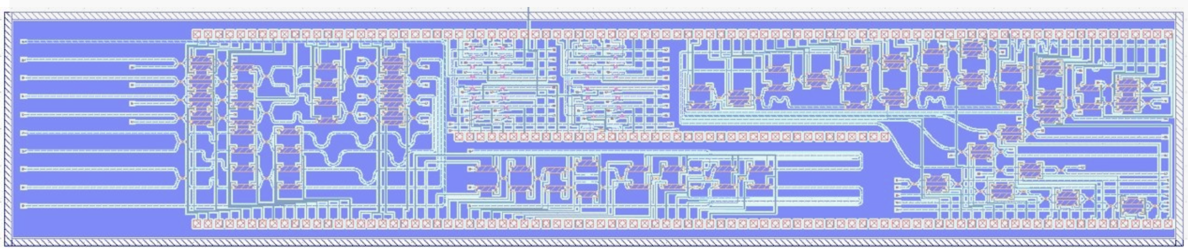


Future Directions

Design for better robustness: FFT non-ideality; weight-encoding error

On-chip training framework for FFT-based ONN architecture

Chip tapeout and experimental testing



SYNOPSYS PHOTONIC SOLUTIONS

amf ADVANCED MICRO FOUNDRY

Thank you !
Q&A