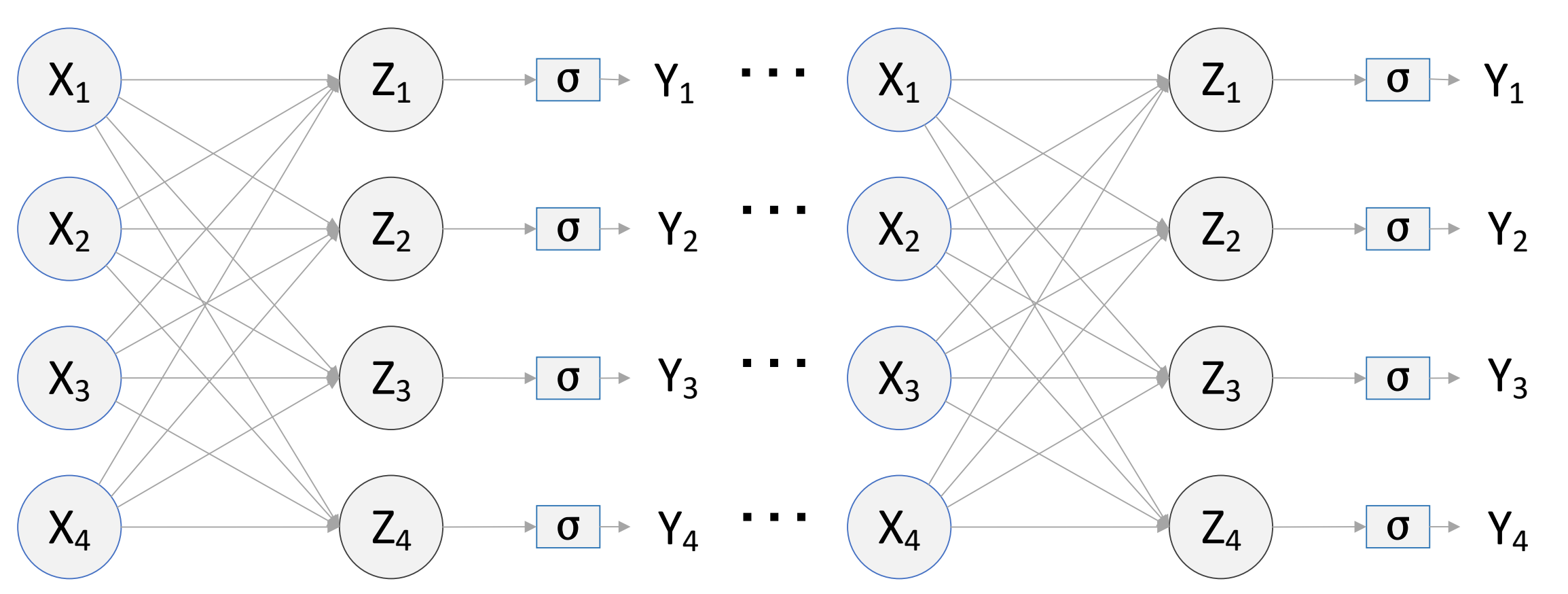
Towards Area-Efficient Optical Neural Networks: An FFT-based Architecture

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

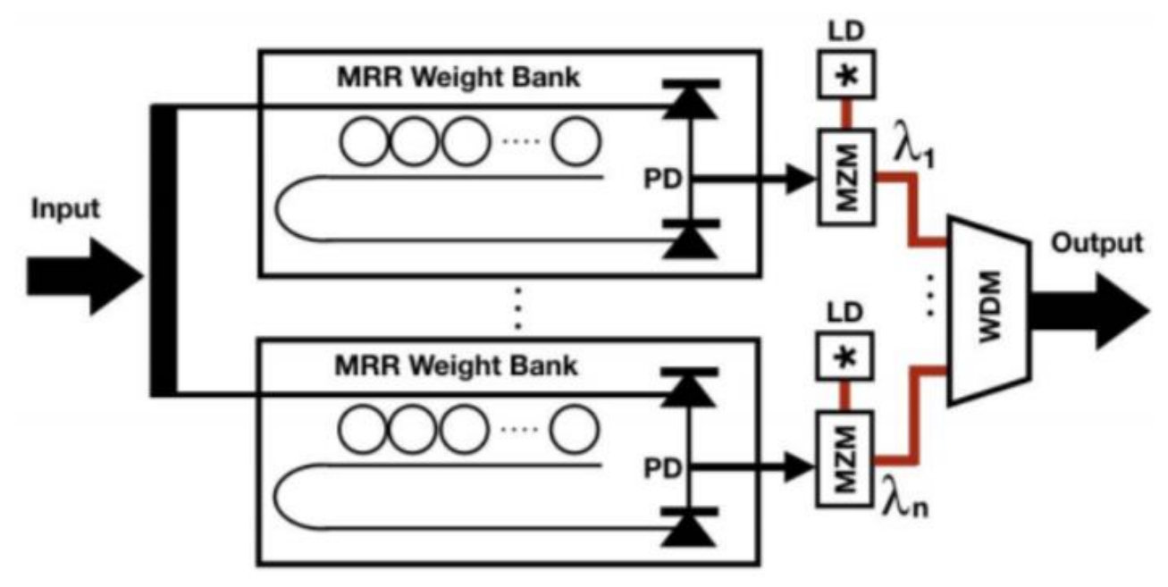
[1]University of Texas at Austin

Multi-layer Perceptron Inference

- Input
  - Vector $\boldsymbol{x}$
- Output
  - Vector $\boldsymbol{y} = \sigma(\boldsymbol{W}\cdot\boldsymbol{x})$
- Objective
  - Accuracy

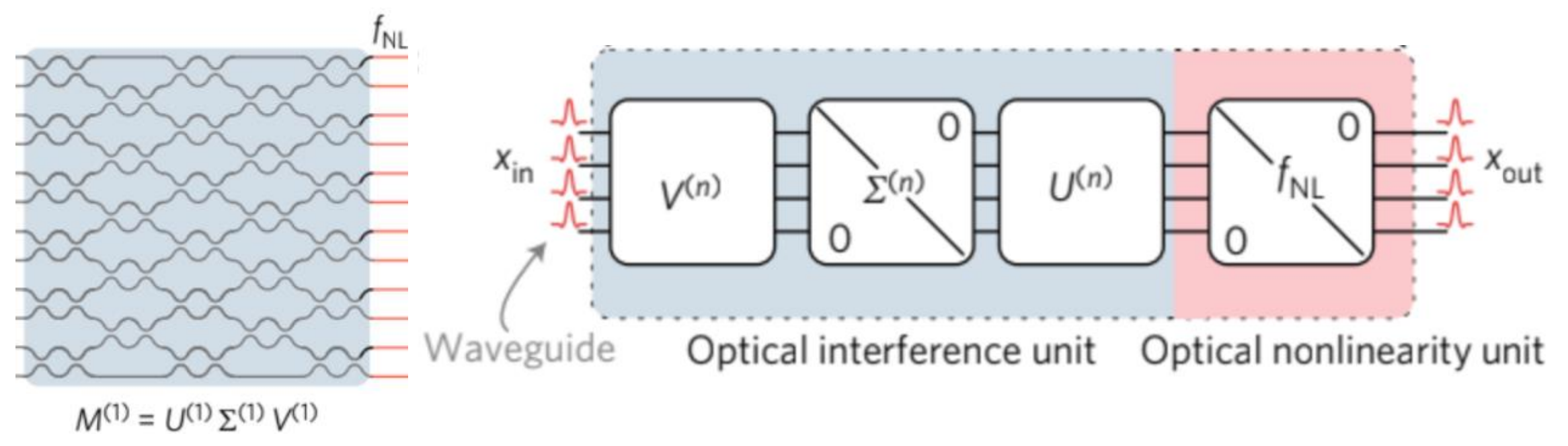


Previous Work on ONNs



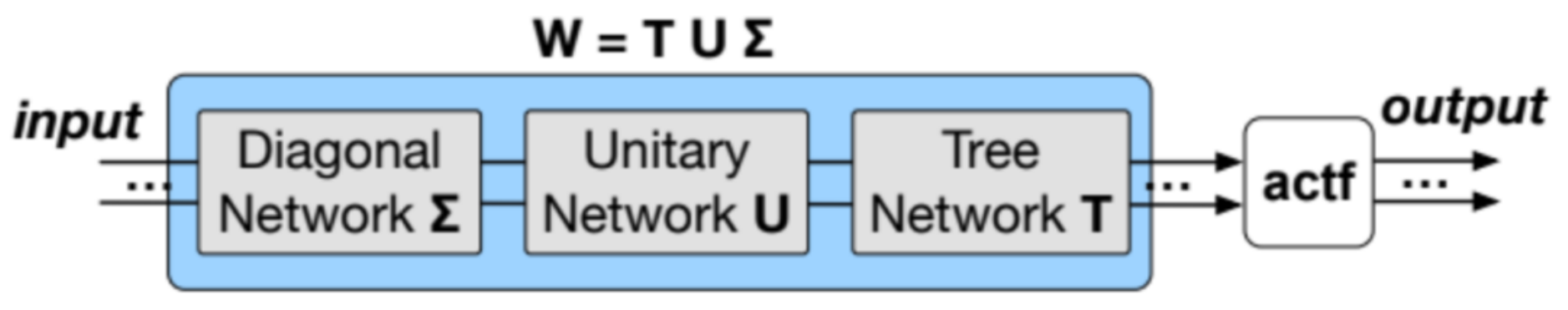
- Photonic microring resonator banks

[IEEE SOCC'18, Mehrabian+]


f_{NL}
x_{in}
$V^{(n)}$
$\Sigma^{(n)}$
$U^{(n)}$
f_{NL}
x_{out}
Waveguide
Optical interference unit
Optical nonlinearity unit
$M^{(i)} = U^{(i)}\Sigma^{(i)}V^{(i)}$

- SVD-based ONNs with MZIs

[Nature'17, Shen+]

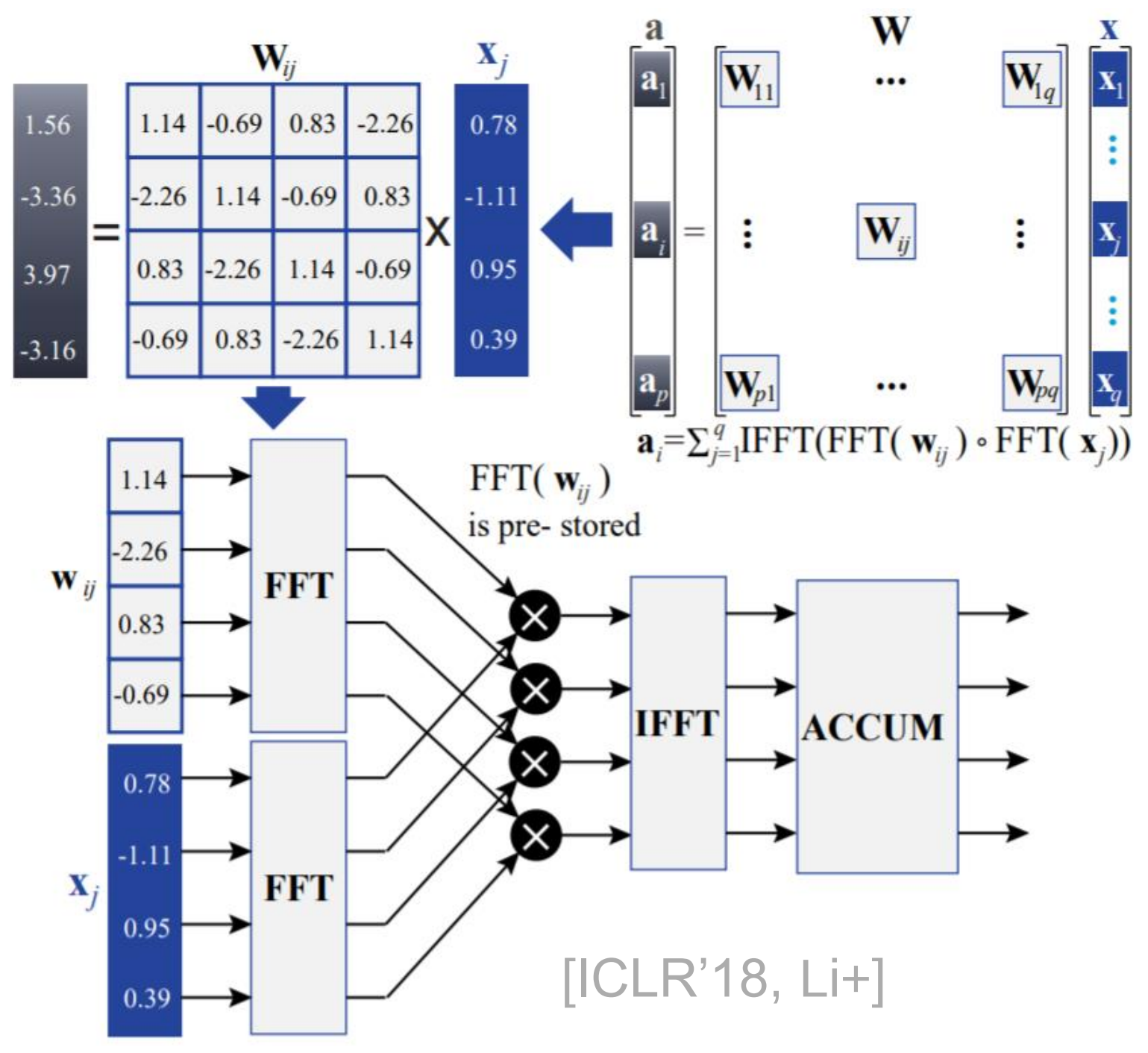


- $T\Sigma U$-based ONNs with MZIs and sparse tree

[ASPDAC'19, Zhao+]

Background on Structured Neural Networks

- **ERNN using SNN** [ICLR'18,Li+]
  - Low computational complexity
  - Low storage complexity
  - ~8x parameter reduction
  - <0.5% accuracy degradation
- **Theoretical Proof** [ICML'17,Zhao+]
  - Universal approximation
  - Identical error bound as classical NNs

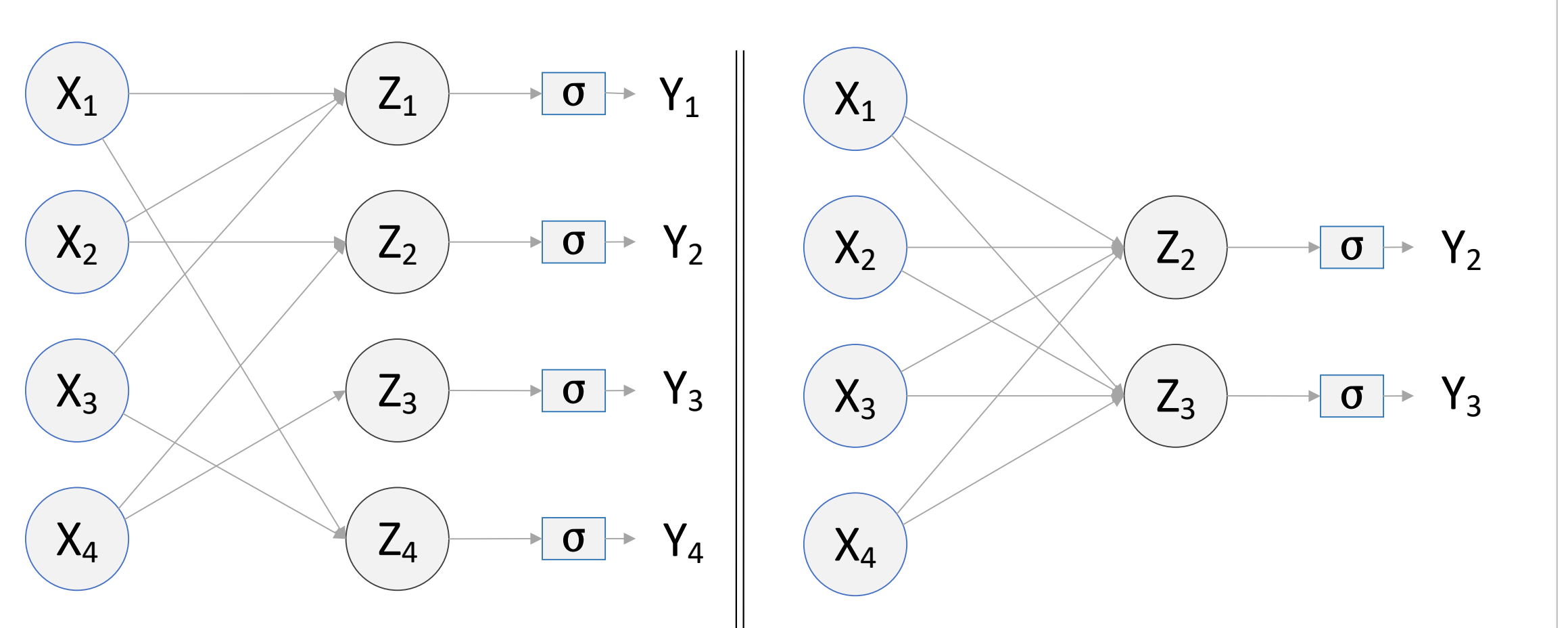


[ICLR'18, Li+]

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} \Longleftrightarrow \boldsymbol{y} = \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{w}) \odot \mathcal{F}(\boldsymbol{x}))$$

Advances in Network Pruning

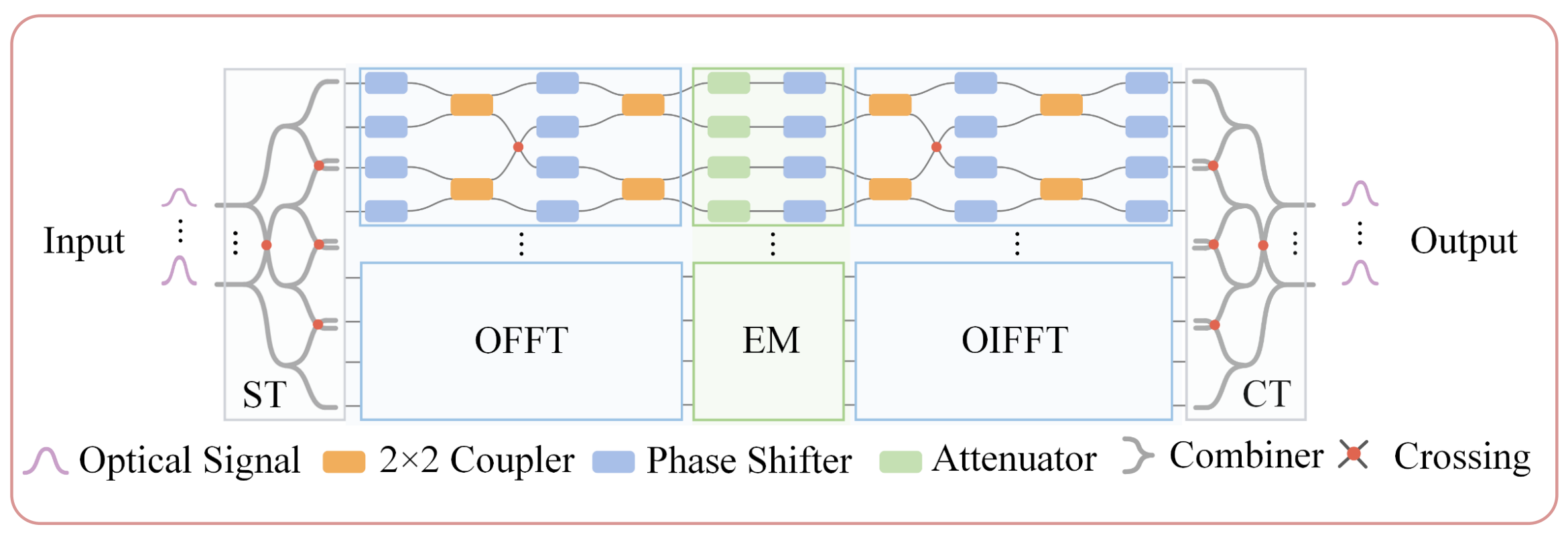
Network slimming with pruning techniques



Non-structured pruning
- Random zero entries
- Irregular

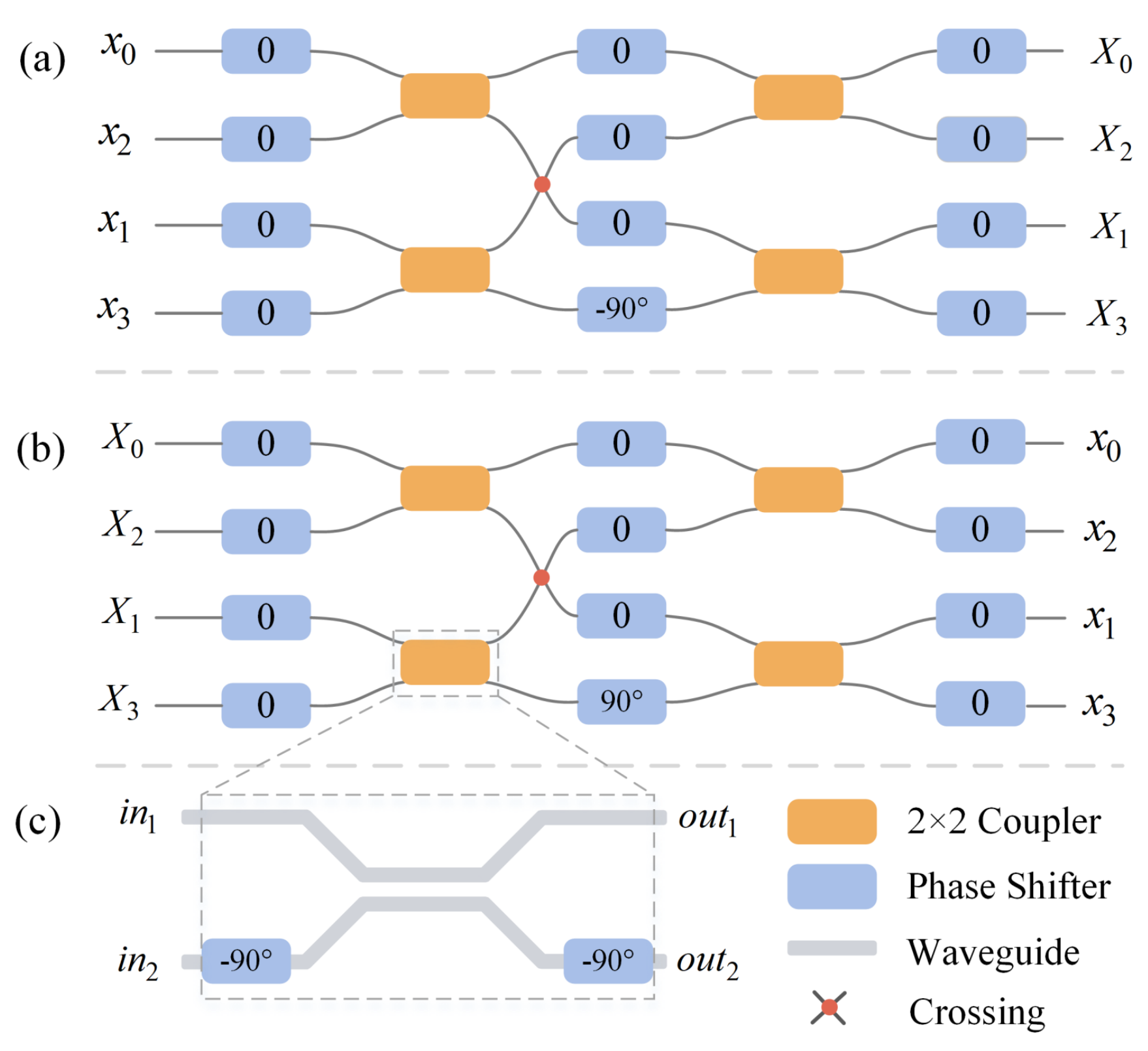
Structured pruning
- Zero entries in group
- Regular
- Hardware-friendly

Proposed ONN Architecture



Optical Fast Fourier Transform

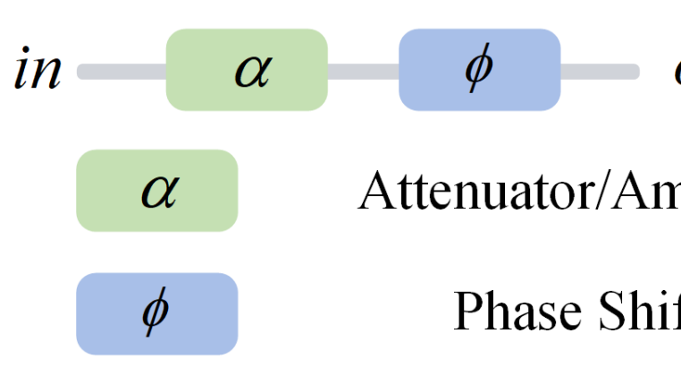
$$\boldsymbol{X}_k = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}\boldsymbol{x}_n e^{-i\frac{2\pi k n}{N}} \quad k = 0, 1, \cdots, N-1.$$



$$\begin{pmatrix}out_1\\ out_2\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}in_1+in_2\\ in_1-in_2\end{pmatrix} = \underbrace{\begin{pmatrix}1&0\\ 0&-j\end{pmatrix}}_{\text{output phase shifter}}\underbrace{\frac{1}{\sqrt{2}}\begin{pmatrix}1&j\\ j&1\end{pmatrix}}_{\text{directional coupler}}\underbrace{\begin{pmatrix}1&0\\ 0&-j\end{pmatrix}}_{\text{input phase shifter}}\begin{pmatrix}in_1\\ in_2\end{pmatrix}$$

- 2 × 2 couplers and phase shifters to achieve OFFT

Element-wise Vector Multiplication

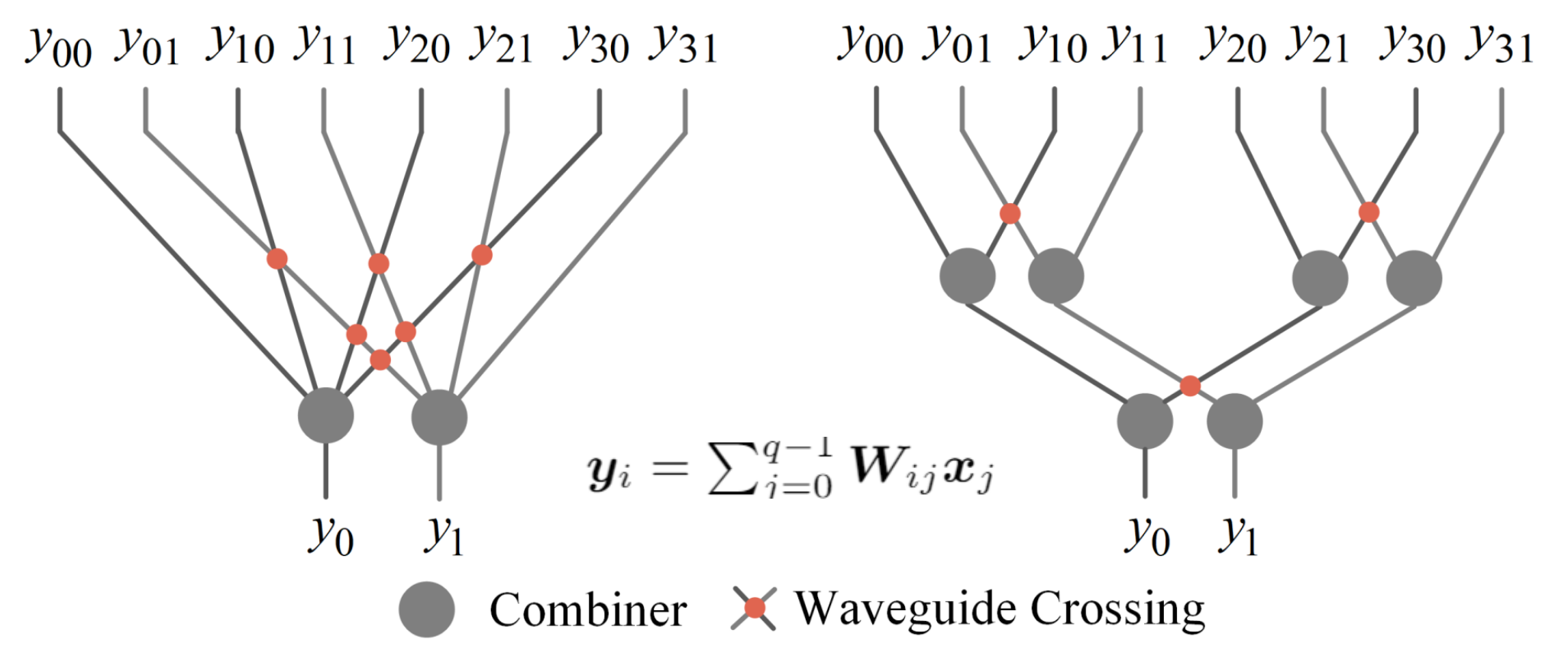


Attenuator/Amplifier

Phase Shifter

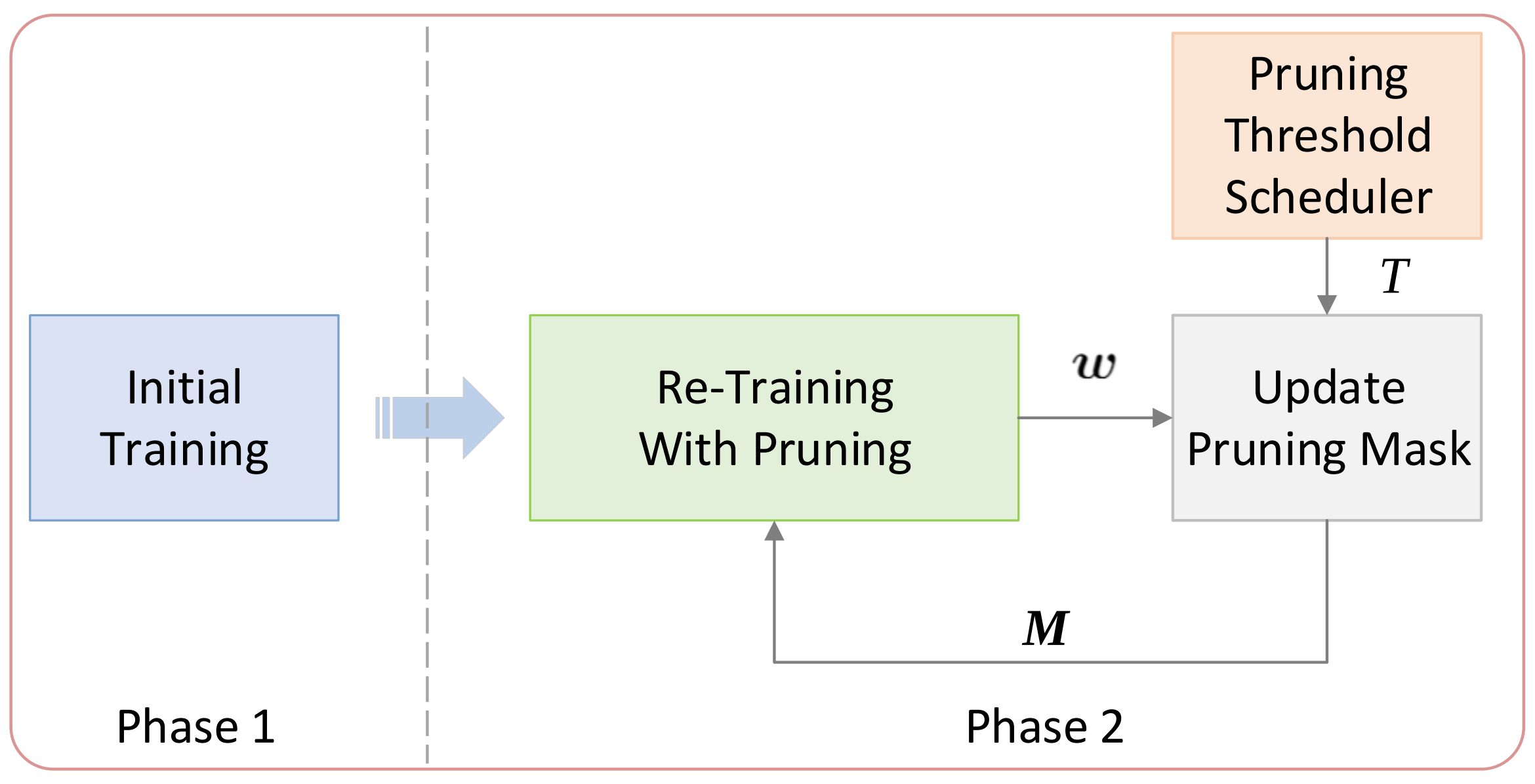
- Complex-valued multiplication
- Light polarization
- Encode weights in attenuators

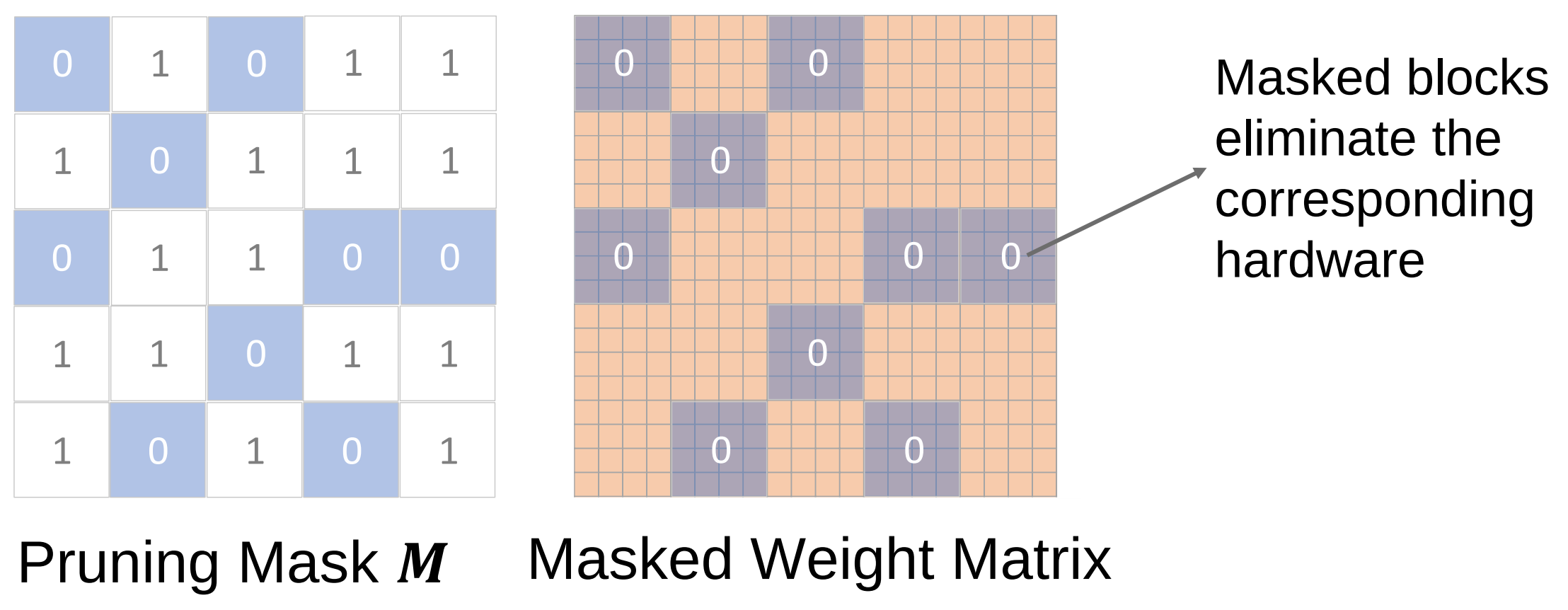
Splitter/Combiner Tree


y_{00} y_{01} y_{10} y_{11} y_{20} y_{21} y_{30} y_{31}
y_0 y_1
$y_i = \sum_{i=0}^{q-1} W_{ij} x_j$
Combiner
Waveguide Crossing

- Fewer waveguide crossings: $k(k-1)(q-1)/2$
- Avoid multi-port combiners

Two-Phase Training Flow with Structured Pruning





Masked blocks eliminate the corresponding hardware

Pruning Mask $\boldsymbol{M}$ Masked Weight Matrix

Hardware Utilization Analysis

SVD-based Architecture $(\mathbf{W} \in \mathbf{R}^{m\times n})$

$$\#DC_{\mathrm{SVD}} = m(m-1) + n(n-1) + \max(m,n)$$
$$\#PS_{\mathrm{SVD}} = \frac{m(m+1)}{2} + \frac{n(n-1)}{2}$$

TΣU-based Architecture $(\mathbf{W} \in \mathbf{R}^{m\times n})$

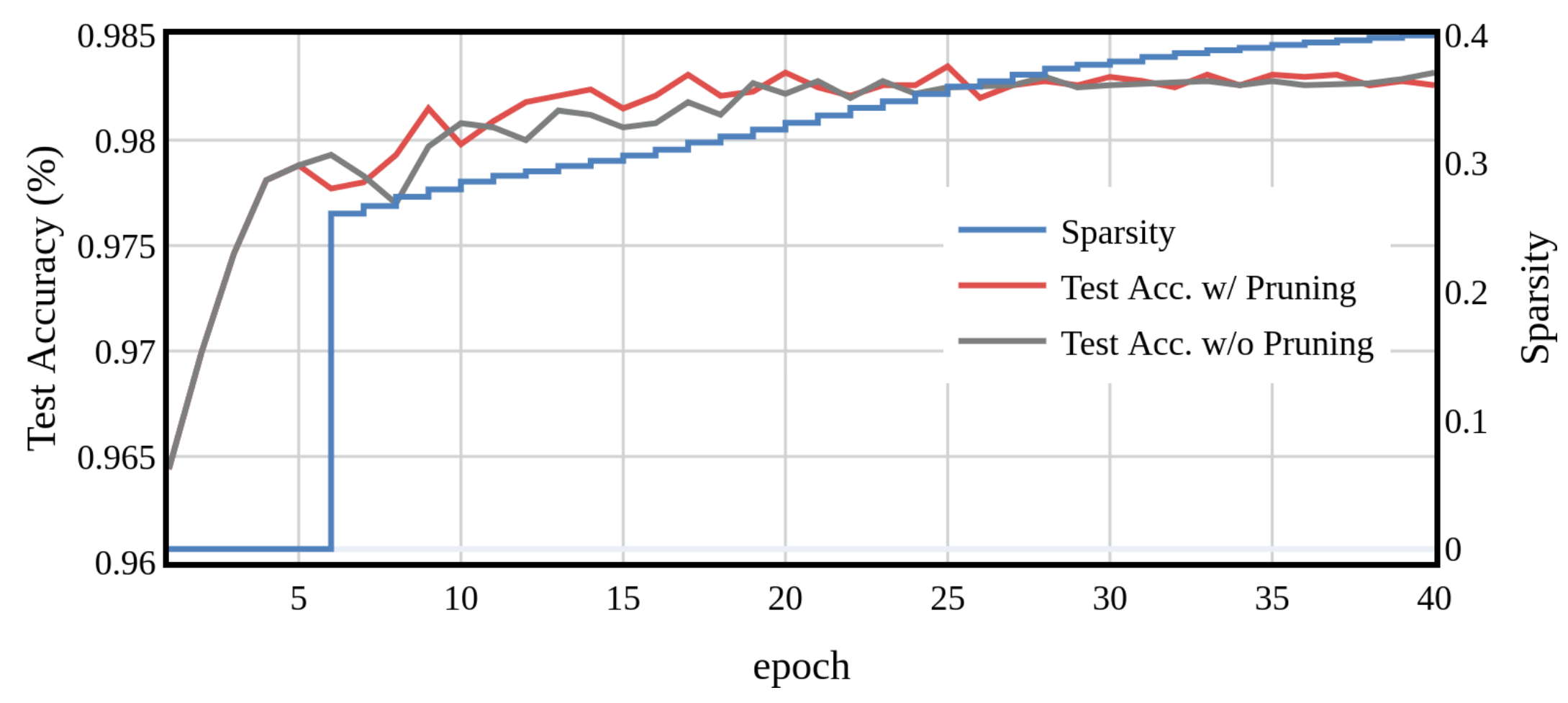
$$\#DC_{\mathrm{T\Sigma U}} = n(n+1) + \max(m,n)$$
$$\#PS_{\mathrm{T\Sigma U}} = \frac{n(n+1)}{2}$$

Ours Architecture $(\mathbf{W} \in \mathbf{R}^{m\times n}, \boldsymbol{block} = \boldsymbol{k})$

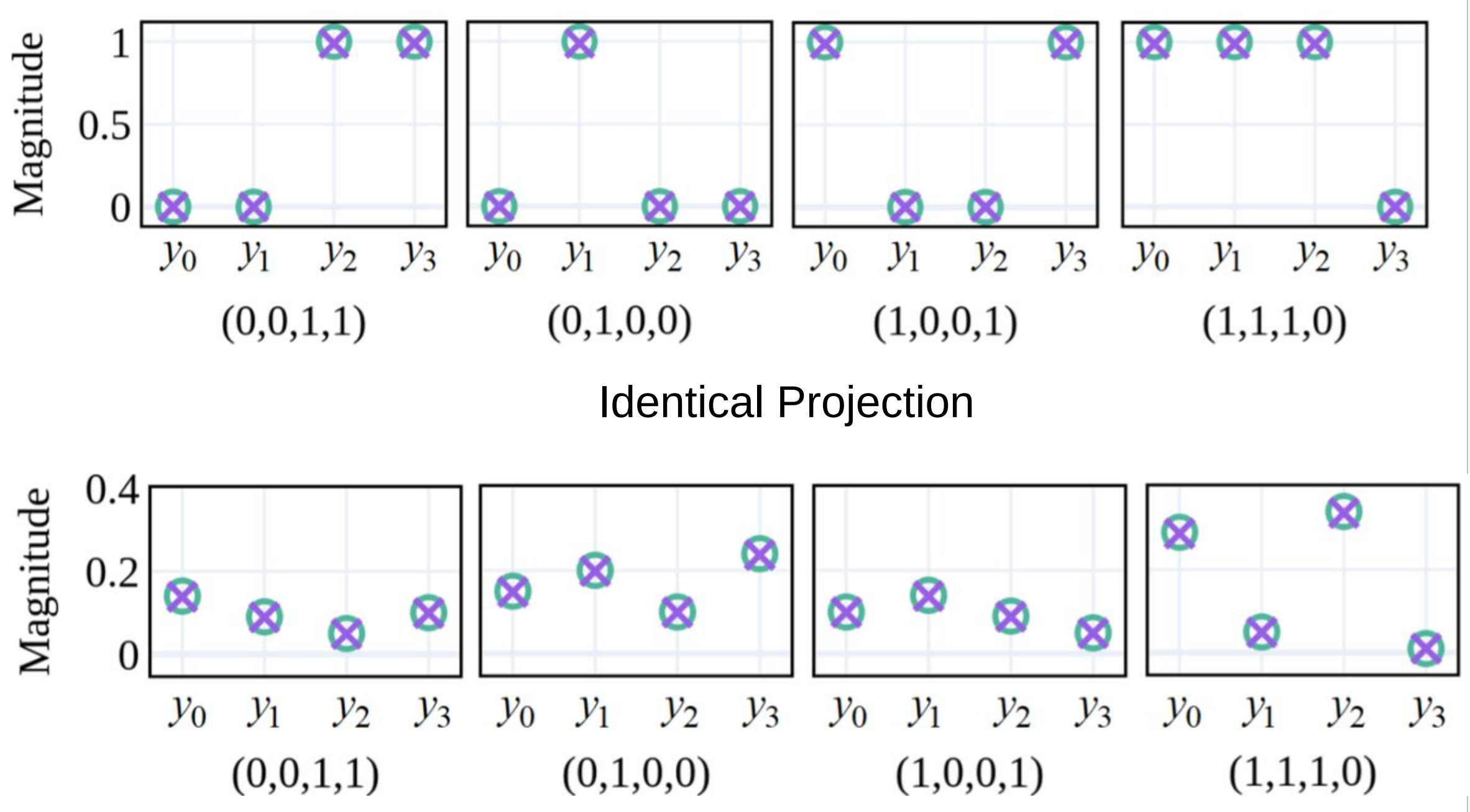
$$\#DC_{\mathrm{Ours}} = \frac{mn}{k}(\log_2 k + 1)$$
$$\#PS_{\mathrm{Ours}} = \frac{mn}{k}(2\log_2 k + 1)$$

Experimental Results

Training Curve

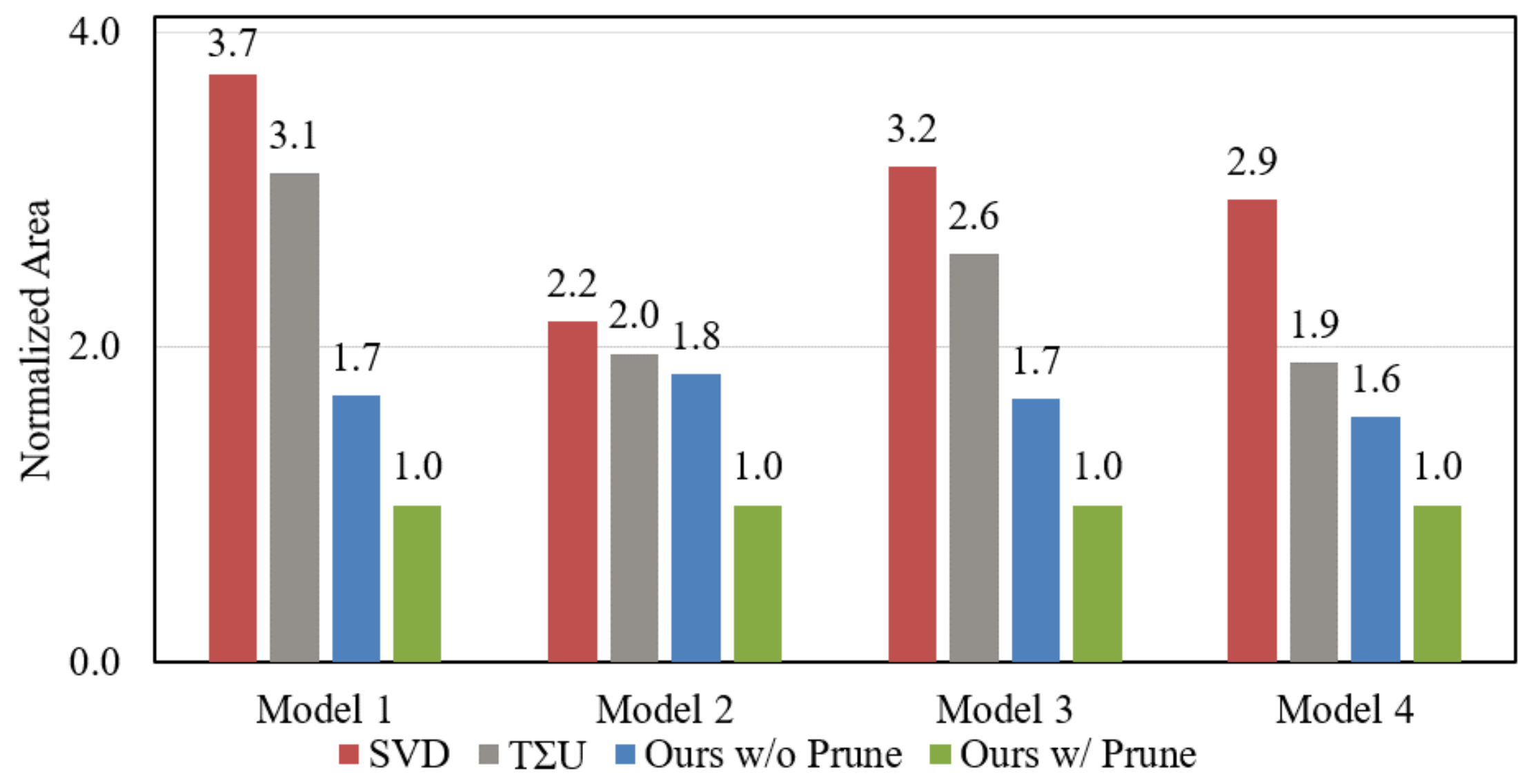


Numerical Simulation (Lumerical)



$$\boldsymbol{w} = (0.2, -0.1, 0.24, -0.15)$$

Area Comparison



Architecture innovation
- Use circulant matrix representation for better efficiency
- Avoid using MZIs
- Friendly to modern network pruning techniques.
- Fewer parameters

Software innovation
- End-to-end training flow to perform structured pruning based on Group Lasso regularization
- Incremental method avoid accuracy degradation

Conclusion and Future Work

- New architecture to save optical component for better area efficiency
- Enable structured pruning to optical neural networks for network slimming without accuracy degradation
- 2.2~3.7x better area cost than SVD-based architecture
- May extend to OCNN and other compact NNs
- Considering more practical hardware information

Source code: https://github.com/JeremieMelo/fft-onn