Optimizing Integrated Photonic Neural Networks under Imperfections

Sanmitra Banerjee

Duke PRATT SCHOOL of ENGINEERING

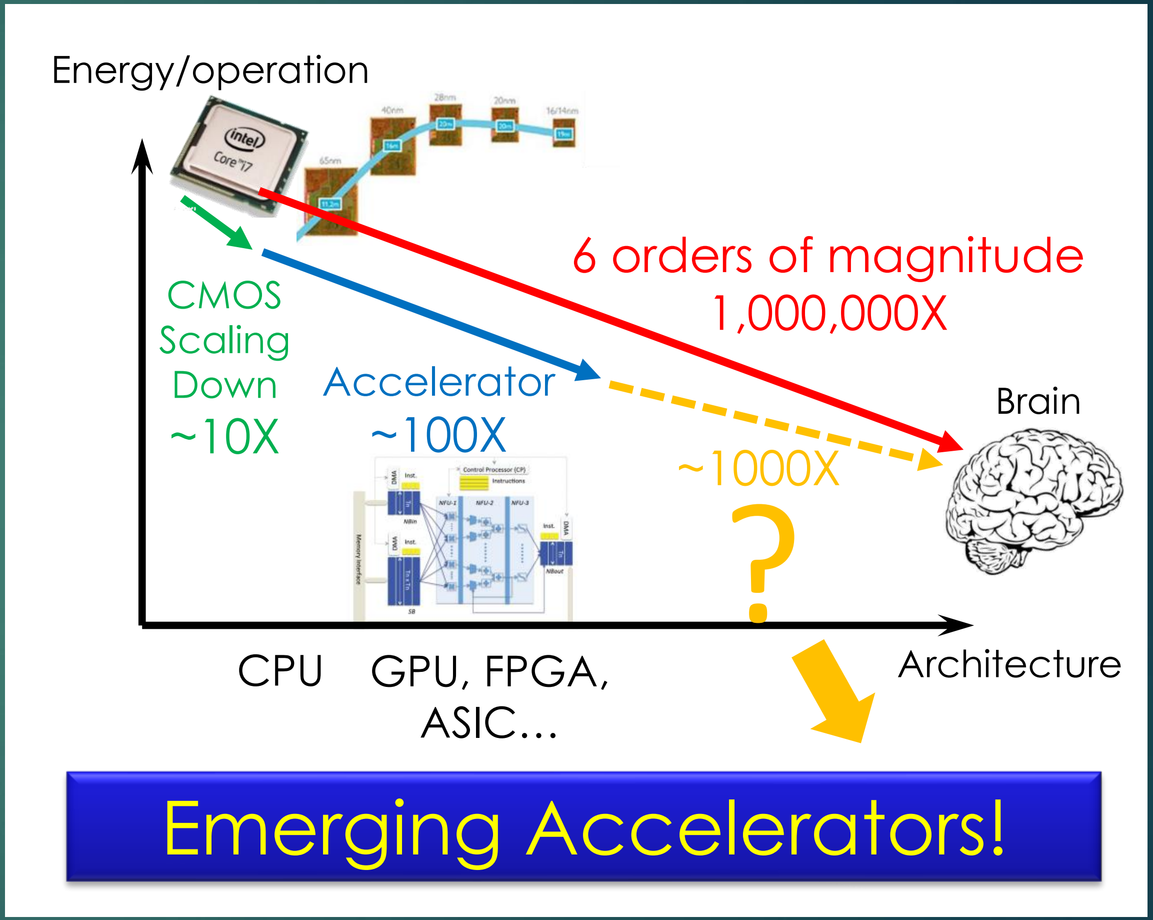
March 01, 2024

Outline

- Background
  - Why integrated photonic Neural Networks (IPNNs)?
  - Sources of imperfections in IPNNs
- Modeling IPNN Imperfections
  - Fabrication uncertainties, quantization errors
  - Thermal crosstalk, insertion loss
- Optimizing IPNNs under Imperfections
  - CHAMP, LTH-Prune, HybridPrune

AI Accelerators

- Accelerators – cornerstones of deep-learning
  - Variable precision
  - Optimized matrix multiplication
- Huge energy efficiency gap
  - DianNao: 452 GOPs/W
  - ISAAC: 800 GOPs/W
  - Brain: 500,000 GOPs/W



[Y. Wang et.al, ISCAS,2016 slides]

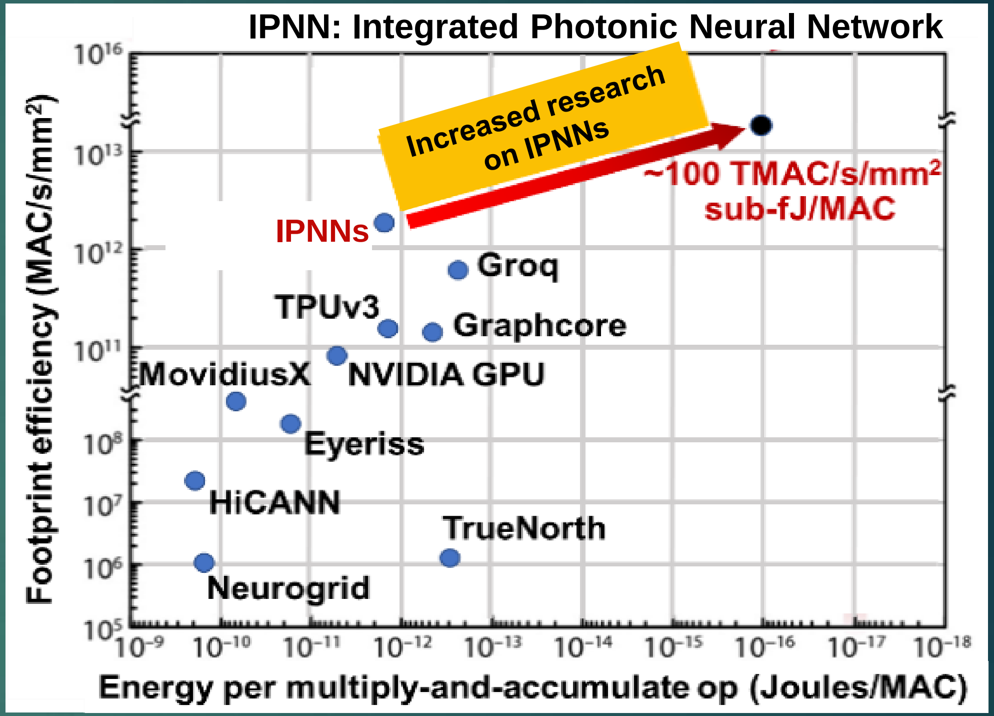
Silicon Photonics

- Computation in optical domain
  - Inherently parallel, at light speed
  - Matrix multiplication in O(1)

Photonics
- High-speed
- Low-loss
- Bulky
- Expensive fabrication

CMOS
- Reliable fabrication
- Small size
- Low-speed
- Power-inefficient

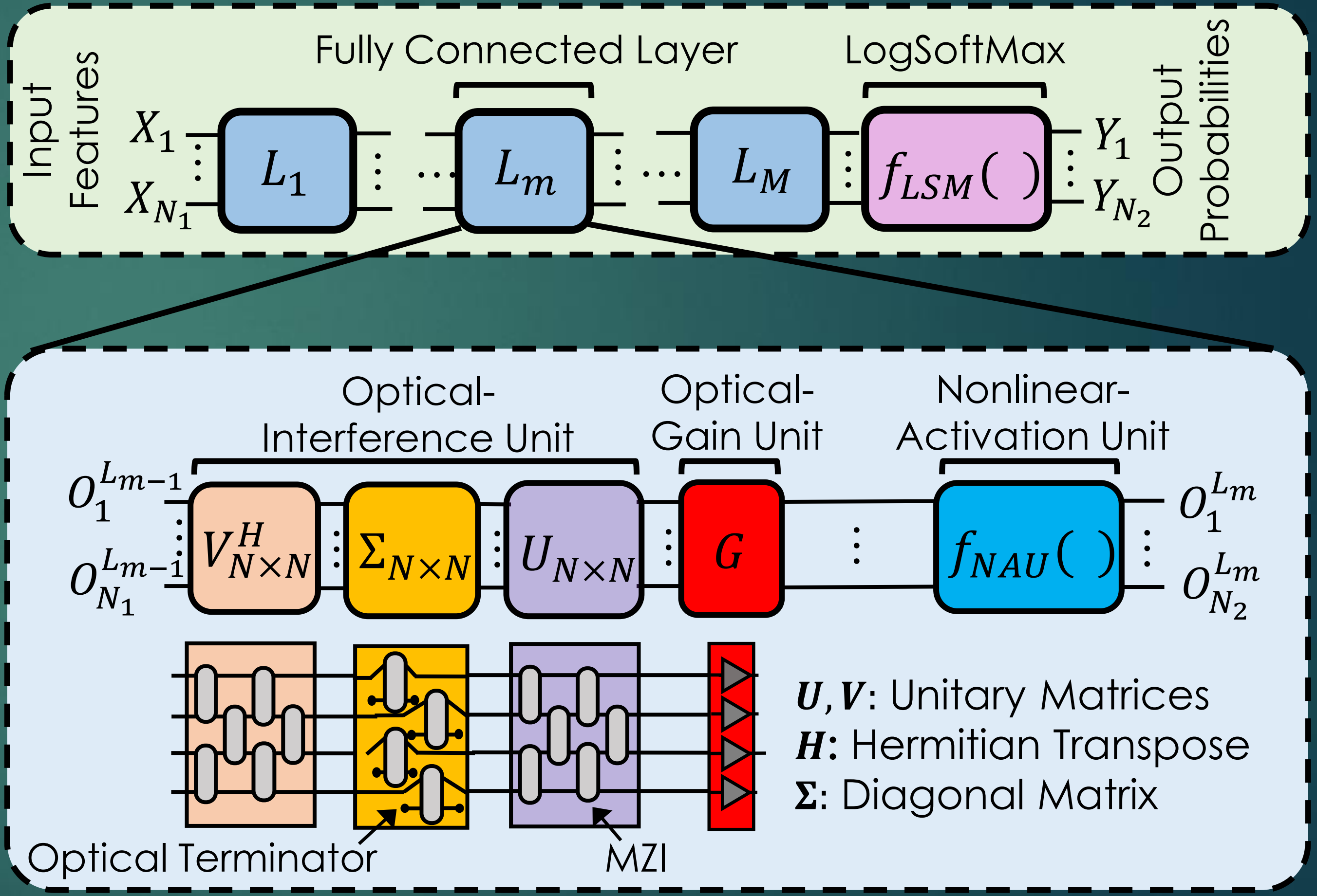
Silicon-photonics
- High-speed
- Low-loss
- Reliable fabrication
- Less bulky



[P.A. Merolla et al., Science, 2014]
[https://web.stanford.edu/group/brainsinsilicon/neurogrid.html]
[https://groq.com/]
[P. Teich, "Tearing apart Google's TPU 3.0 AI coprocessor," 2018]
[A. Reuther et al., HPEC, 2019]

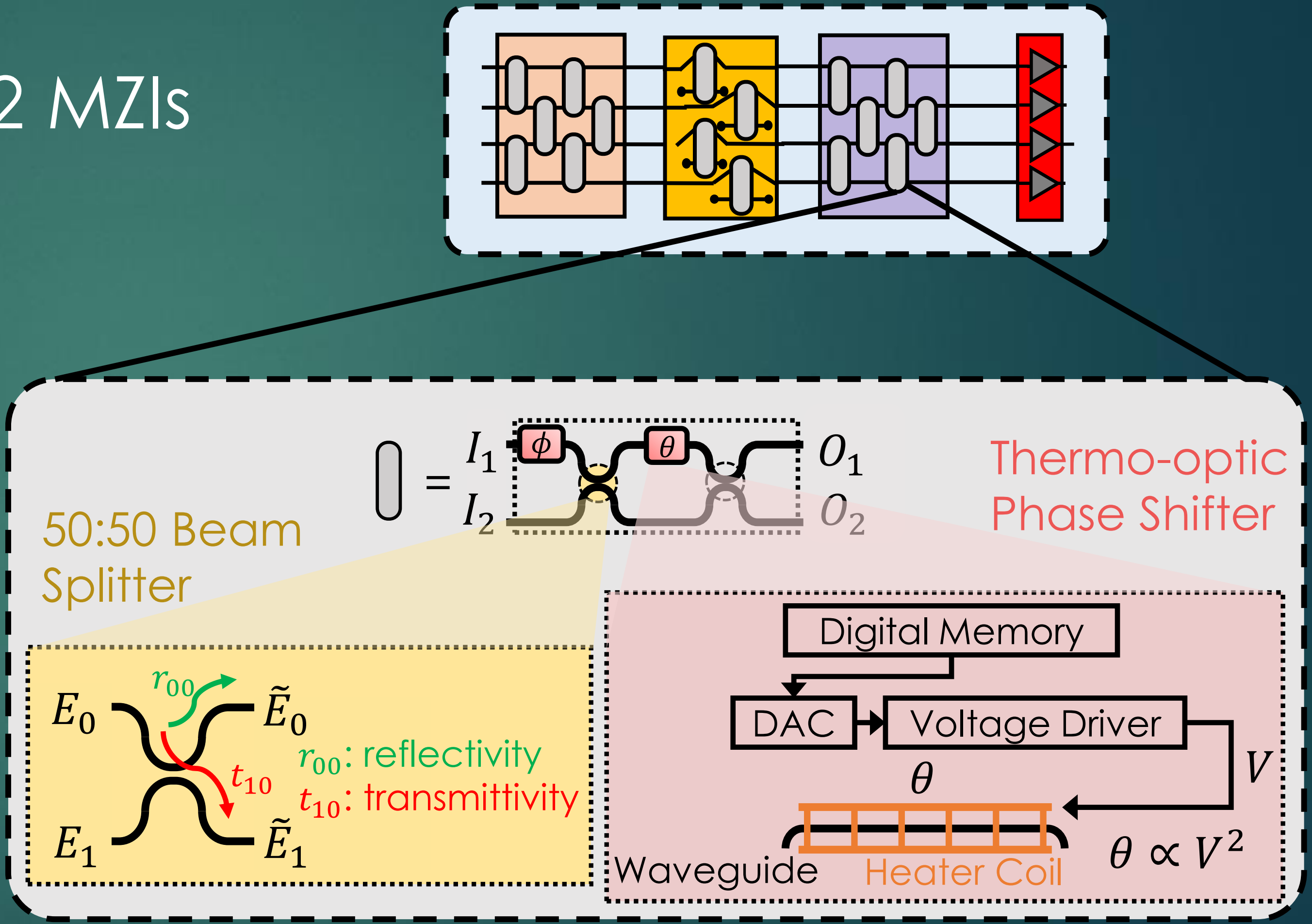
Coherent Photonic NNs

- NNs – cascaded multipliers
- Singular value decomposition (SVD)
- Unitary & diagonal transforms
  - Array of Mach-Zehnder interferometers (MZIs)



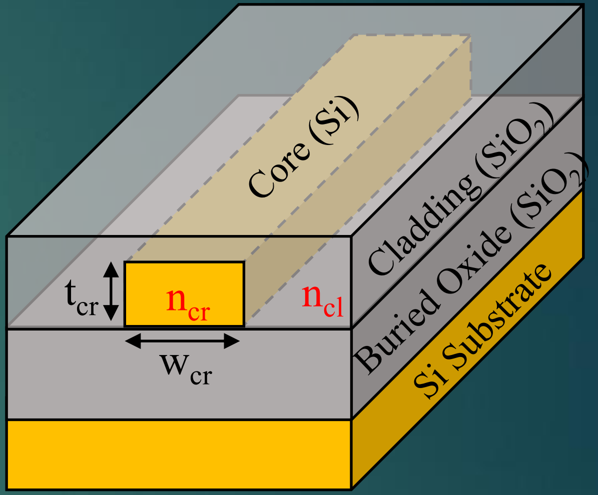
Coherent Photonic NNs

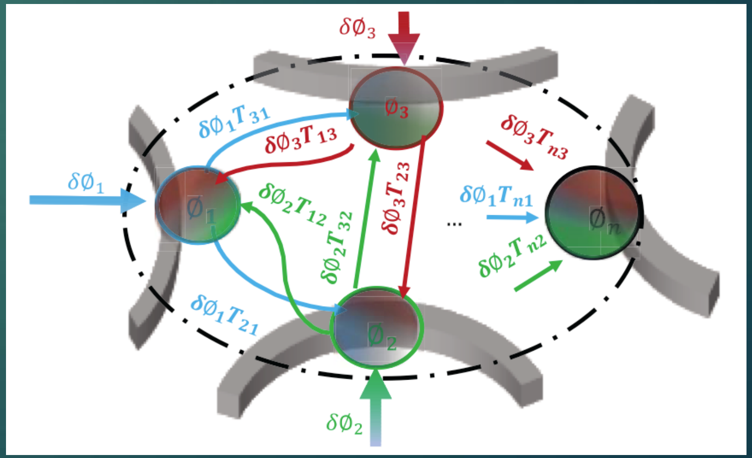
- N×N unitary → N(N-1)/2 MZIs
- MZIs – Phase shifters and beam splitters
- Photonic Training
  - Tune phase angles to minimize loss



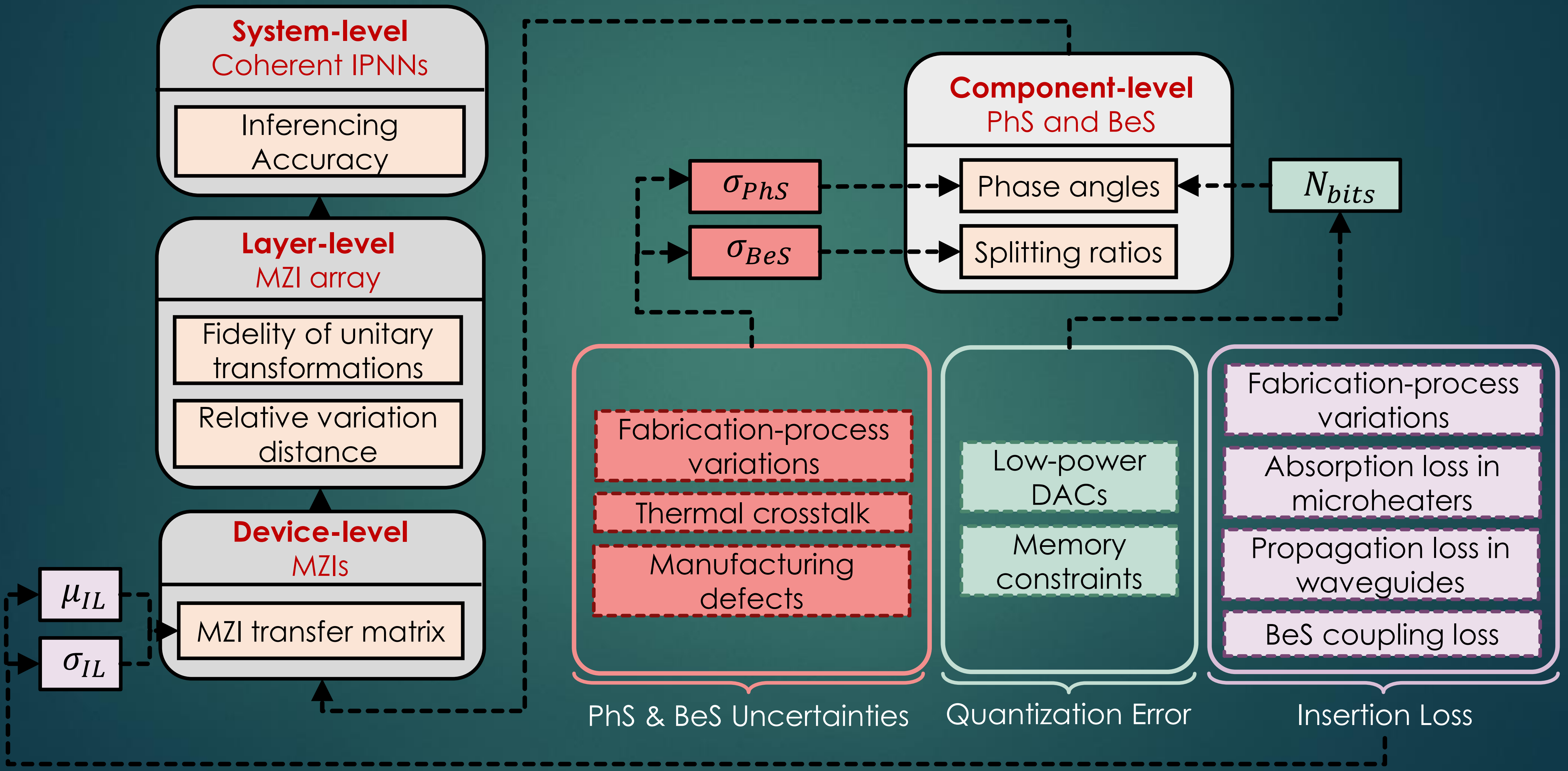
Imperfections in IPNNs

- Nanometer-scale lithographic variations
  - Waveguide width and thickness
  - Length of phase shifters
- Thermal crosstalk
- Non-uniform MZI insertion loss
- Low precision phase encoding – quantization error





Bottom-up Modeling Framework



Uncertainties in PhS and BeS

- Thermo-optic PhS

Lithographic variations

$$\Delta\phi = \frac{2\pi L}{\lambda_0}\frac{dn}{dT}\Delta T$$

Thermal crosstalk, low-precision drivers

  - Average error of ~0.21 radians expected

- 50:50 BeS
  - 1-2% deviation from 50:50 splitting ratio

Component-level
PhS and BeS

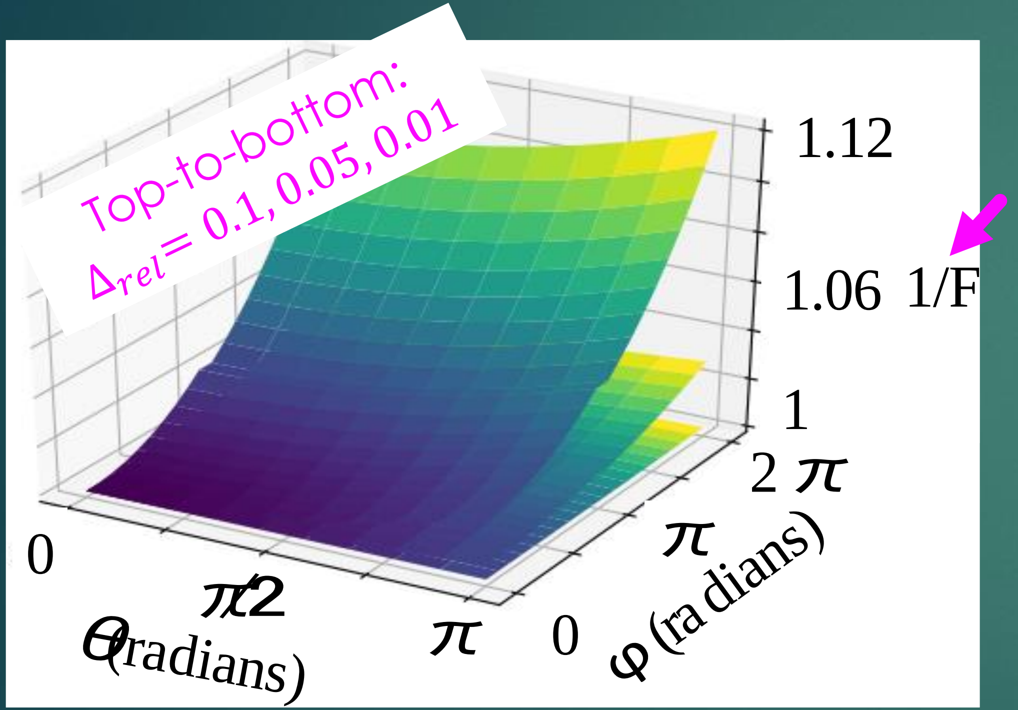
Phase angles

Splitting ratios

[A Crespi, Nature Photonics, 2013]

Uncertainties in PhS and BeS

- Fidelity: closeness between transfer matrices



Device-level
MZIs

MZI transfer matrix

$$T_{MZI}(\theta,\phi) = ie^{i\theta/2}\begin{bmatrix} e^{i\phi}\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \\ e^{i\phi}\cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \end{bmatrix}$$

$$F(T,\tilde{T}) = \left|\frac{Trace(\tilde{T}^{\dagger}T)}{N}\right|^2$$

- Higher phase angles ➔ susceptible to uncertainties

Uncertainties in PhS and BeS

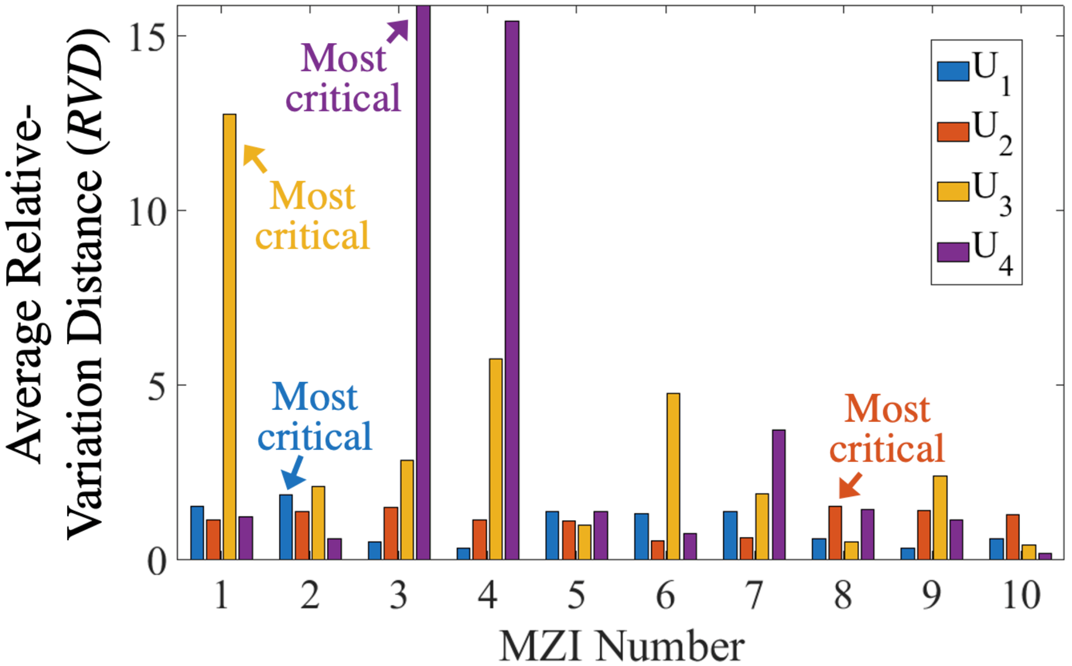
- T_{MZI} deviates ➔ unitary matrix changes

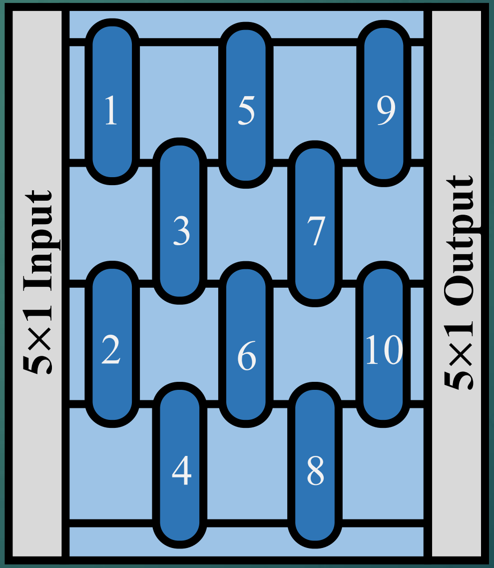
$$RVD(U,\widetilde{U}) = \frac{\sum_m \sum_n |U_{m,n} - \widetilde{U}_{m,n}|}{\sum_m \sum_n |U_{m,n}|}$$

Layer-level
MZI array

Fidelity of unitary transformations

Relative variation distance





$U_{5\times5}$ (Unitary)

$\sigma_{PhS} = \sigma_{BeS} = 0.05$

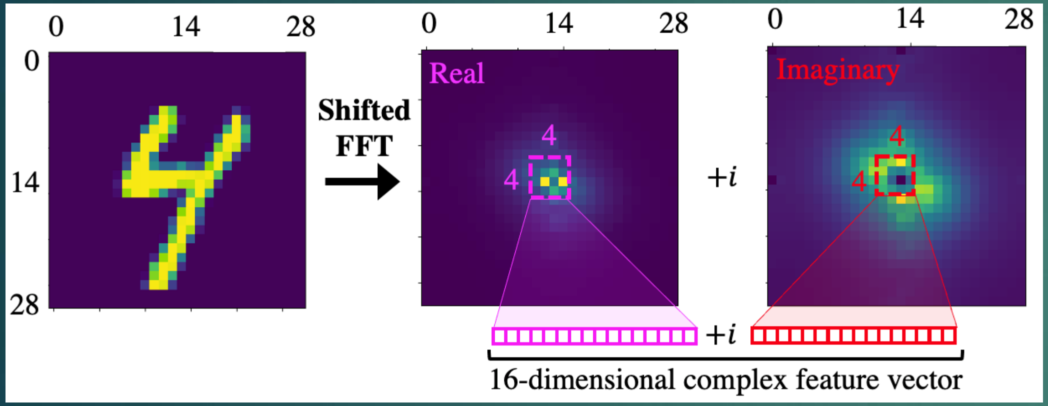
Mean RVD over 1000 iterations

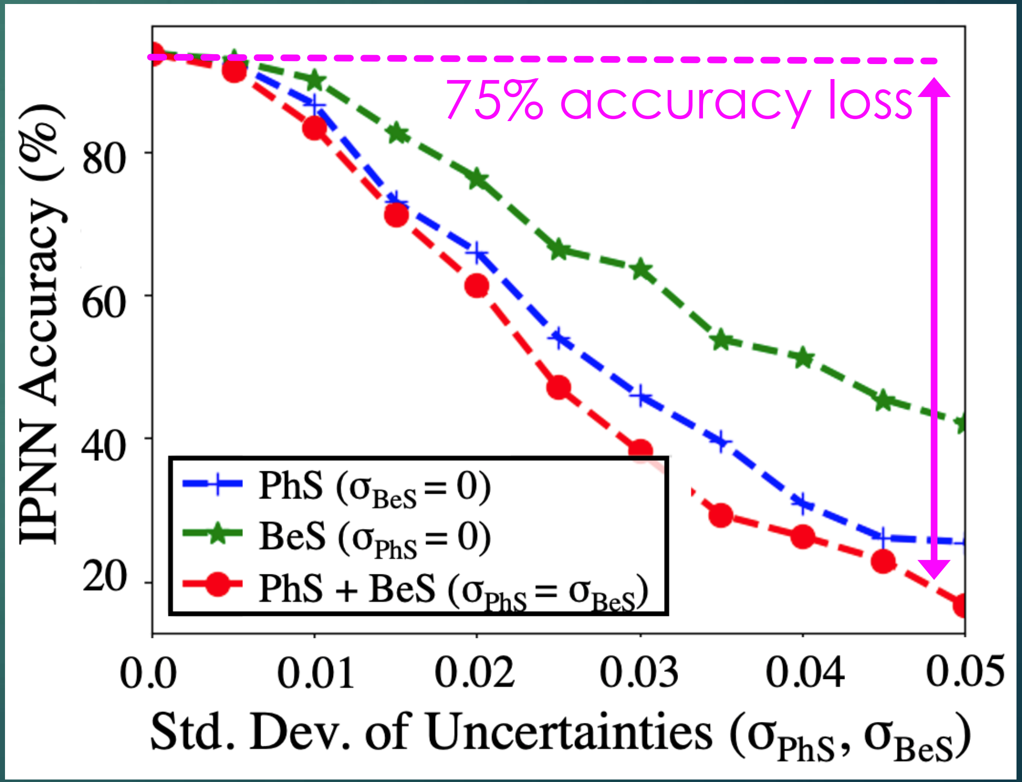
Uncertainties in PhS and BeS

System-level
Coherent IPNNs

Inferencing
Accuracy

- Faulty matrix multiplication – lower accuracy

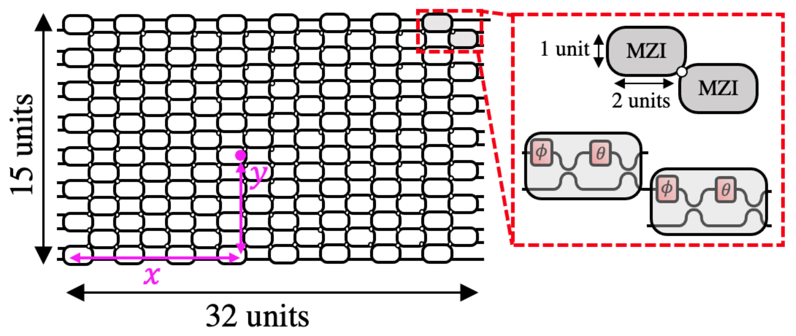


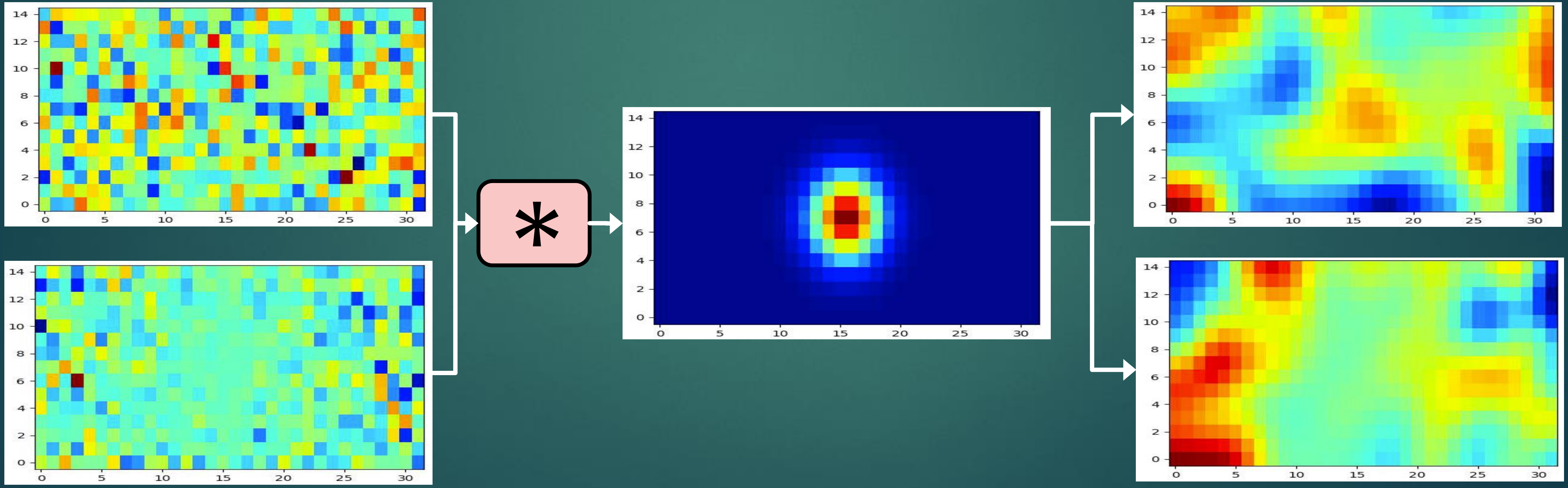


- MLP with two hidden layers
  - 16-16-16-10 (3 multipliers)
  - Nom. accuracy = 93.86%

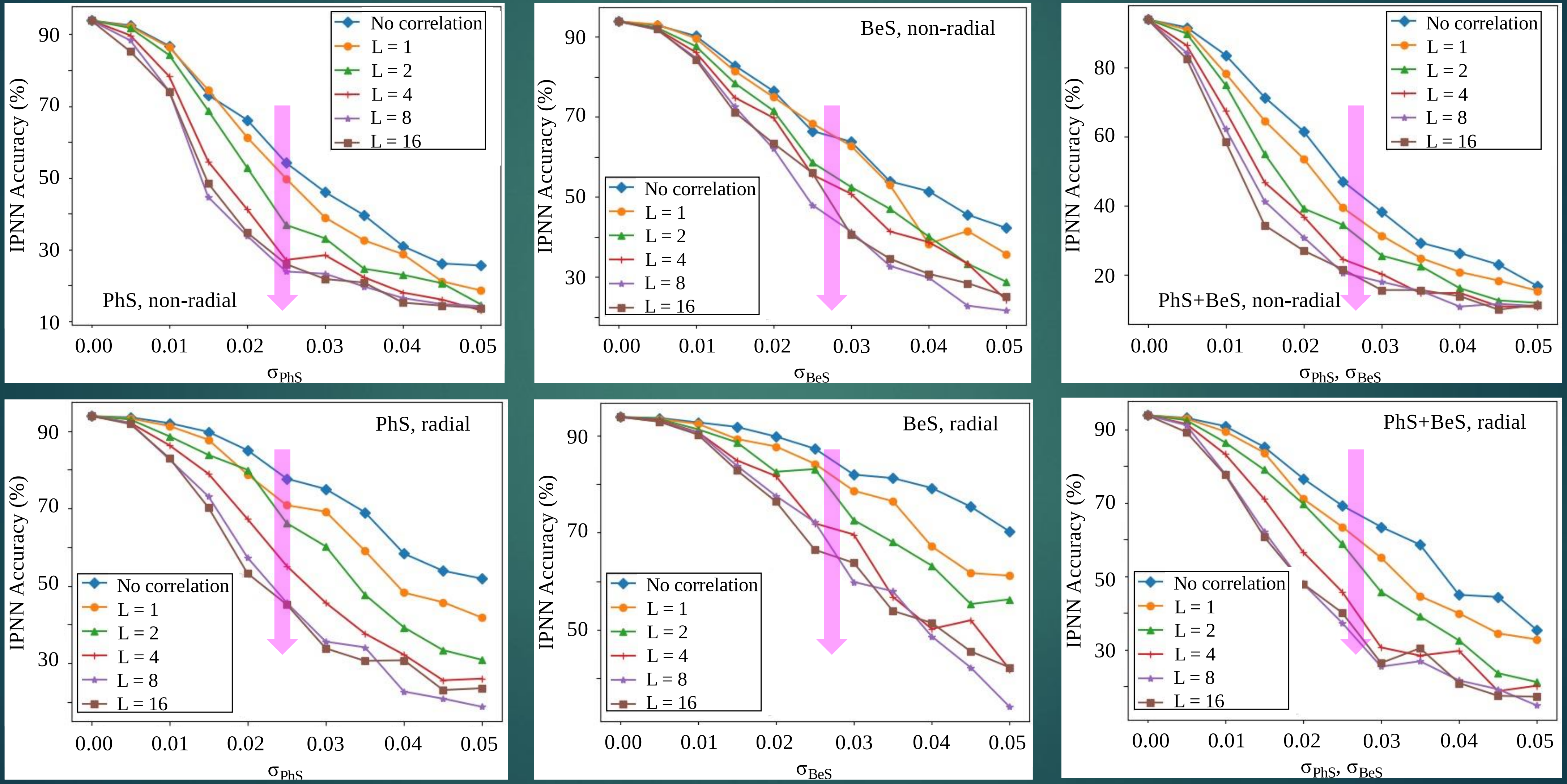
Uncertainties in PhS and BeS

- Photonic uncertainties can be spatially correlated





Uncertainties in PhS and BeS



Non-Uniform MZI Insertion Loss

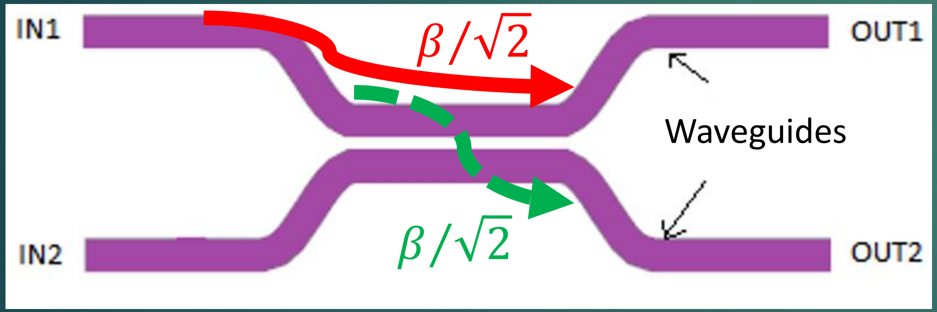
- MZIs are lossy devices
  - Non-uniform loss due to variations

$$IL = 10\log\beta^4$$

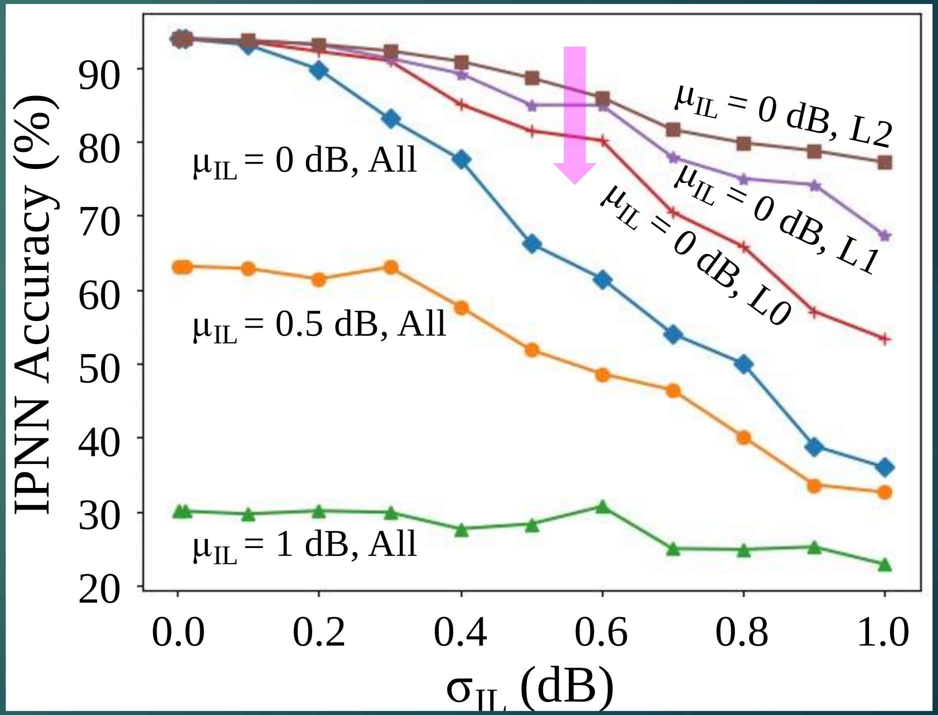
$$IL = \mu_{IL} + ℵ(0, \sigma_{IL}^2)$$

Coupling Loss

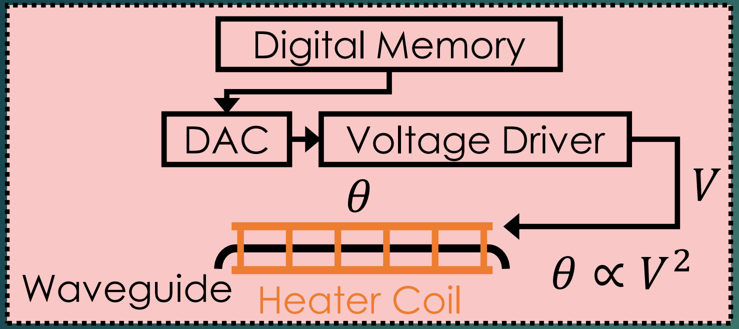
Absorption in microheaters



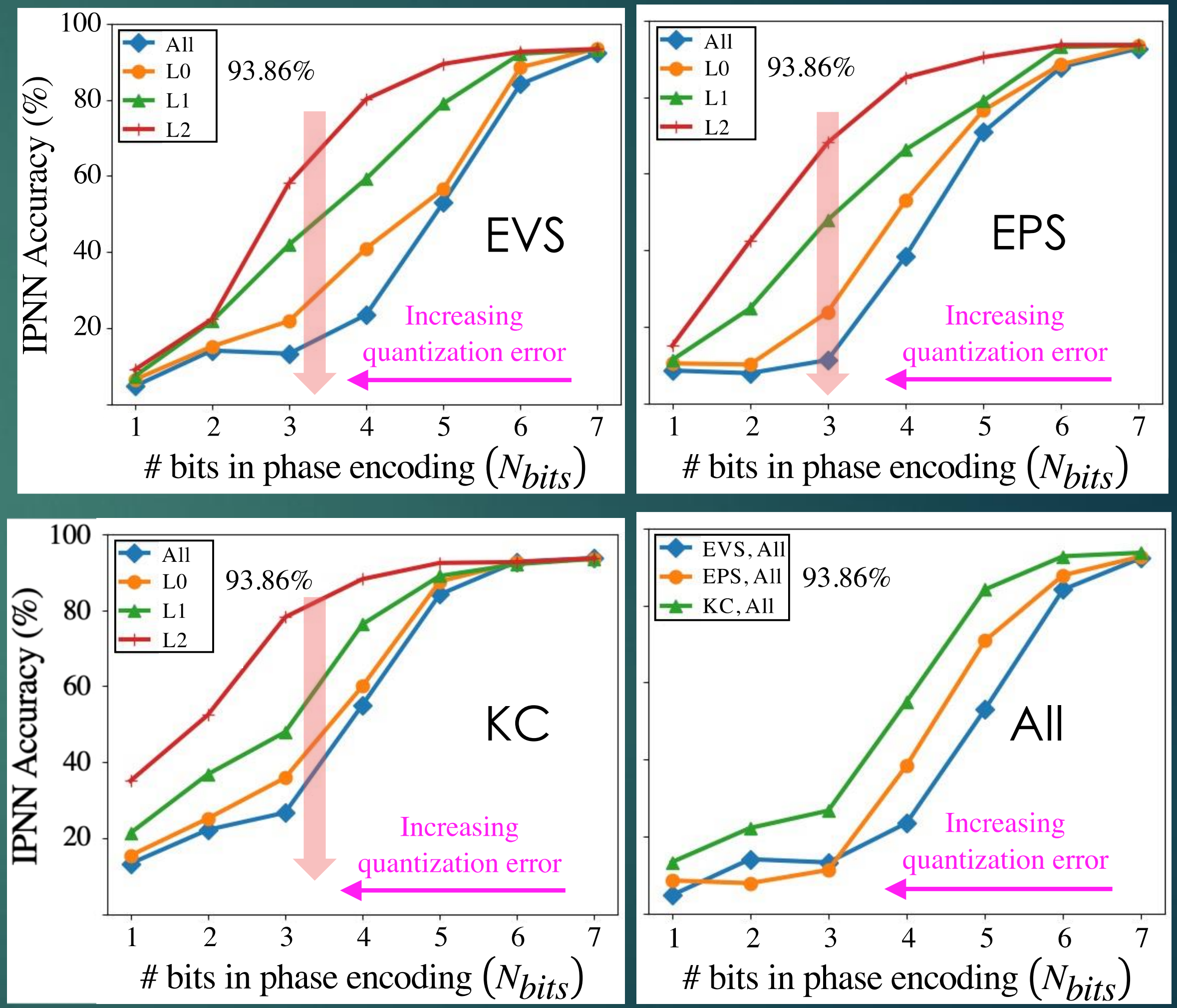
Waveguide propagation



Low-Precision Phase Encoding



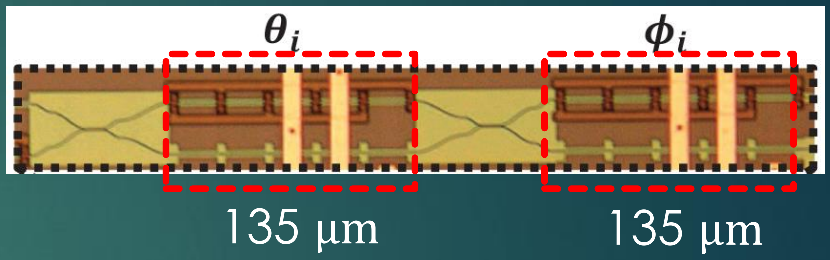
- Memory ↓, DAC power ↓
- Quantization Error ↑
- Equidistant Voltage Steps (EVS)
- Equidistant Phase Steps (EPS)
- K-Means Clustering (KC)



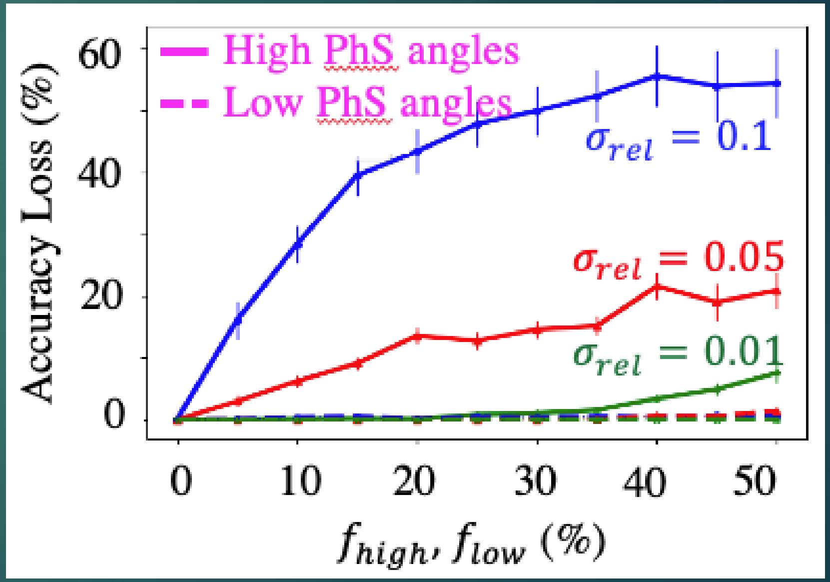
Why Prune Photonic NNs?

- Pruning NNs – reduce parameters with minimal accuracy loss
- Phase Shifters have large footprint
  - O(N^2) phase shifters for N bits data

Pruning phase shifters is essential to the scalability of photonic NNs!



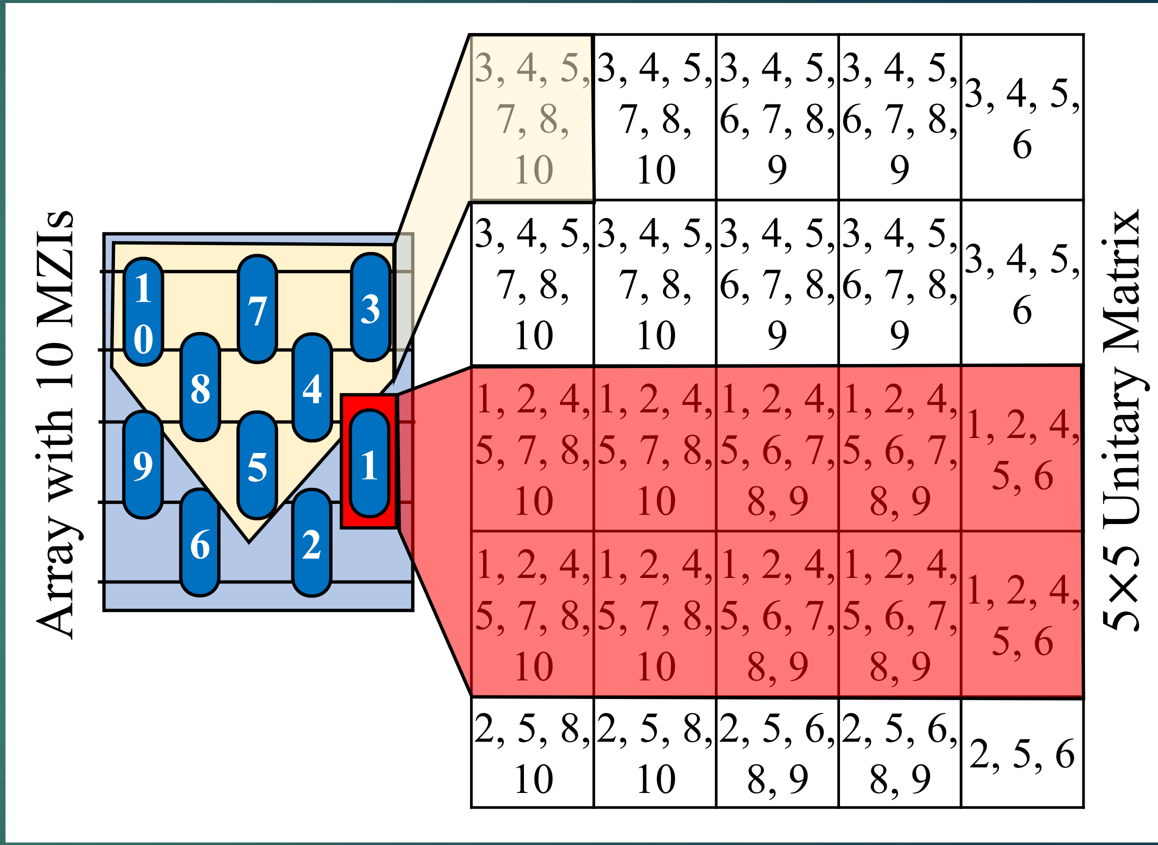
[F. Shokraneh et al., JLT, 2021]



[S. Banerjee et al., OFC, 2021]

Challenges

- DNN pruning: clamp small weights ➔ retrain ➔ sparse weight matrix
- Sparse weight matrix ≠ sparse phase angles
- Bidirectional many-to-one mapping between weights and phase shifters



Software pruning of weight matrices does NOT reduce overhead

Hardware-Aware Pruning: CHAMP

[S. Banerjee et al., *OFC*, 2022]

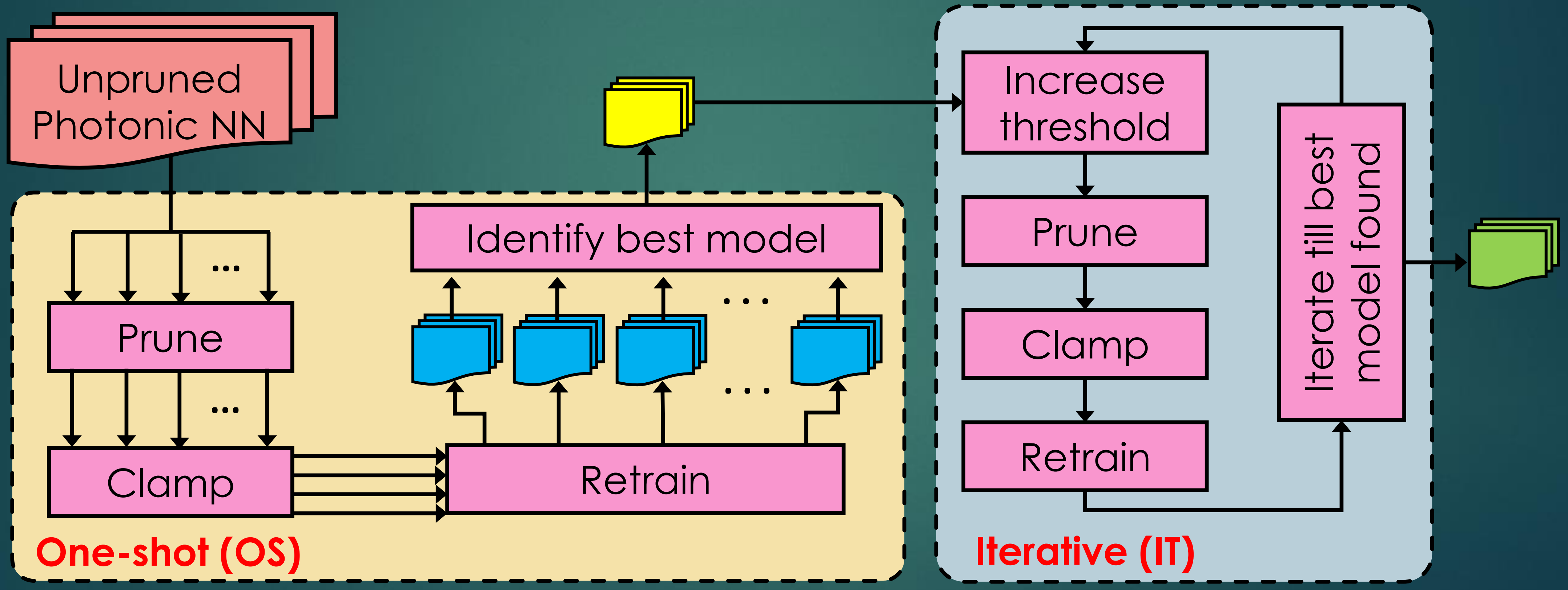
- Pruning Aim: Sparse phase angles, not sparse weights
- Hardware-unaware software pruning does not work
  - Only 30% phase shifters pruned in SOTA

CHAMP: First effective pruning method for photonic NNs

- Photonic training
  - Backpropagation on phase angles, not weights
  - Iteratively clamp phase angles, not weights

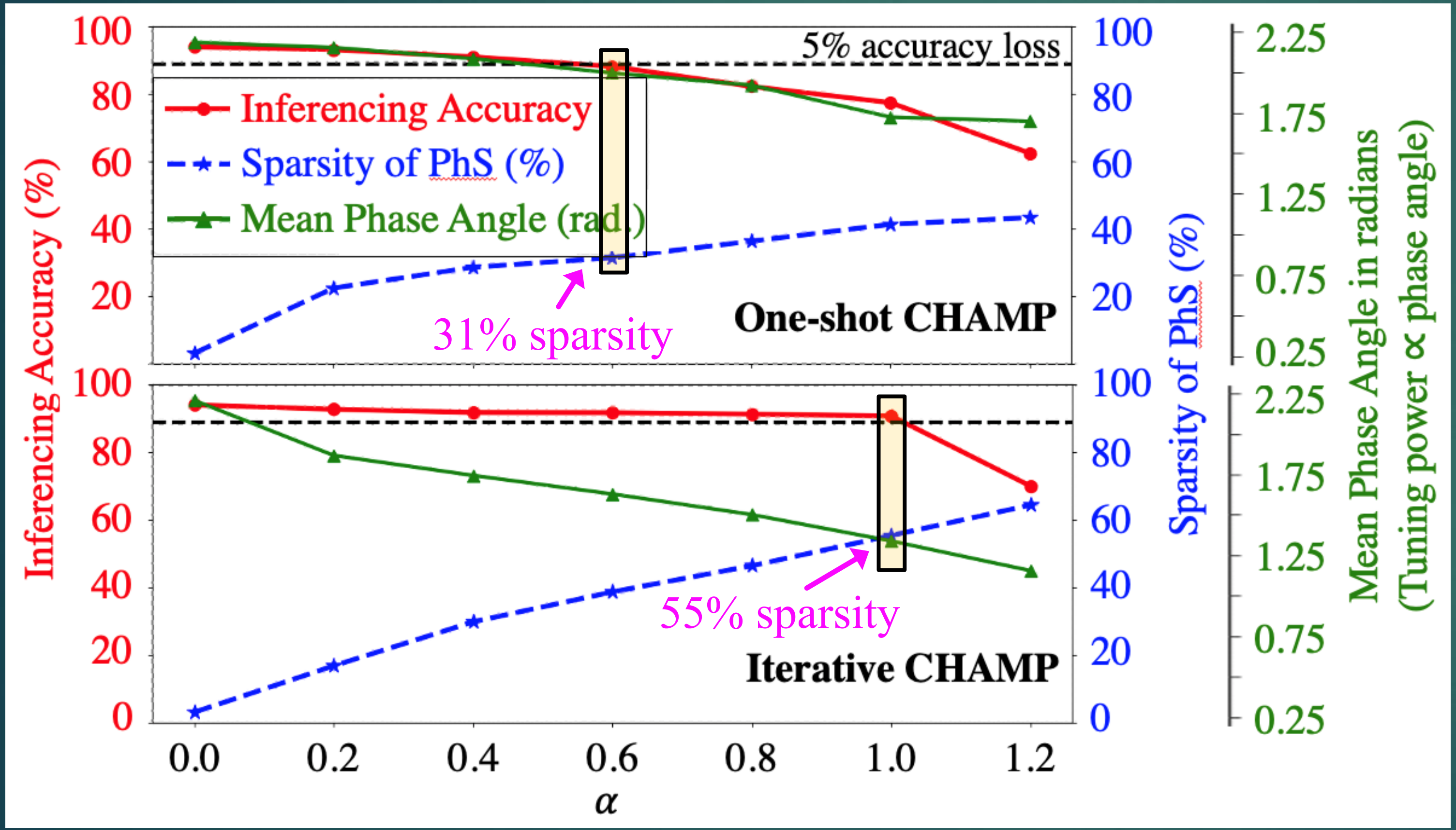
Hardware-Aware Pruning: CHAMP

- Prune phase angles below threshold ➔ clamp ➔ retrain



Simulation Results – CHAMP

- 2 hidden layers with 16 neurons each – 1374 phase angles



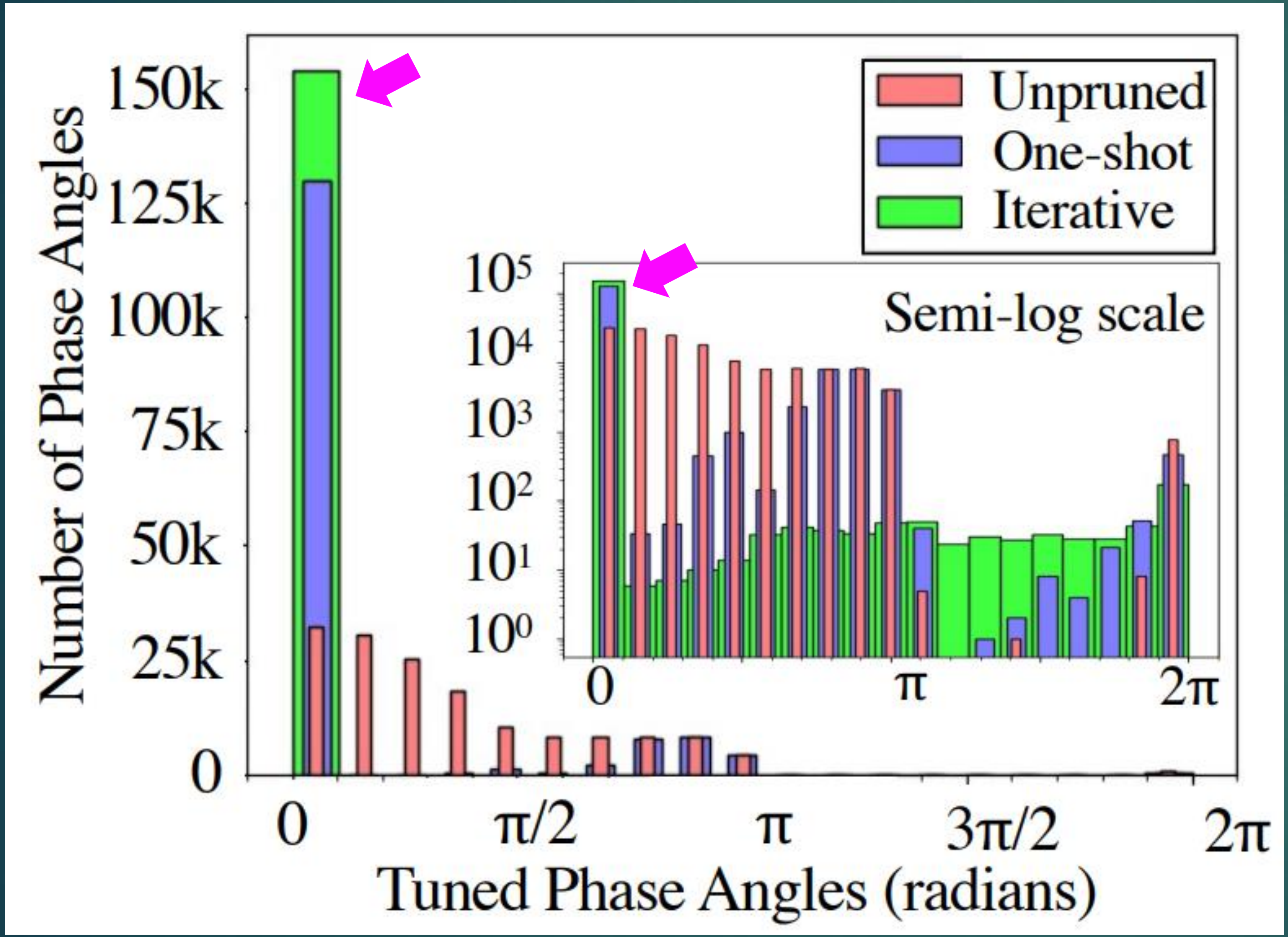
One-Shot Pruning

- Fast
- Parallelized

Iterative Pruning

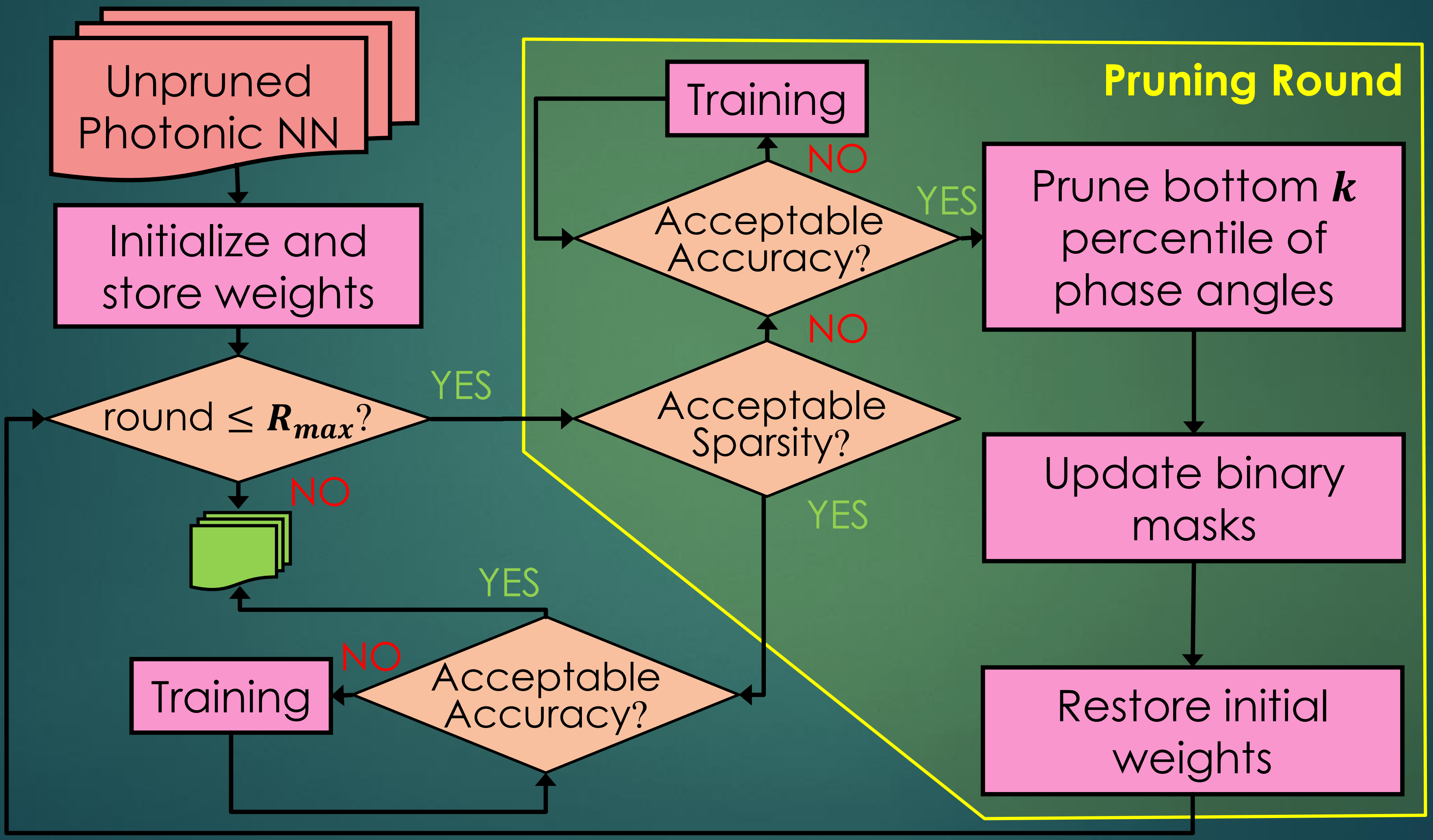
- Gradual
- Low accuracy loss

Simulation Results

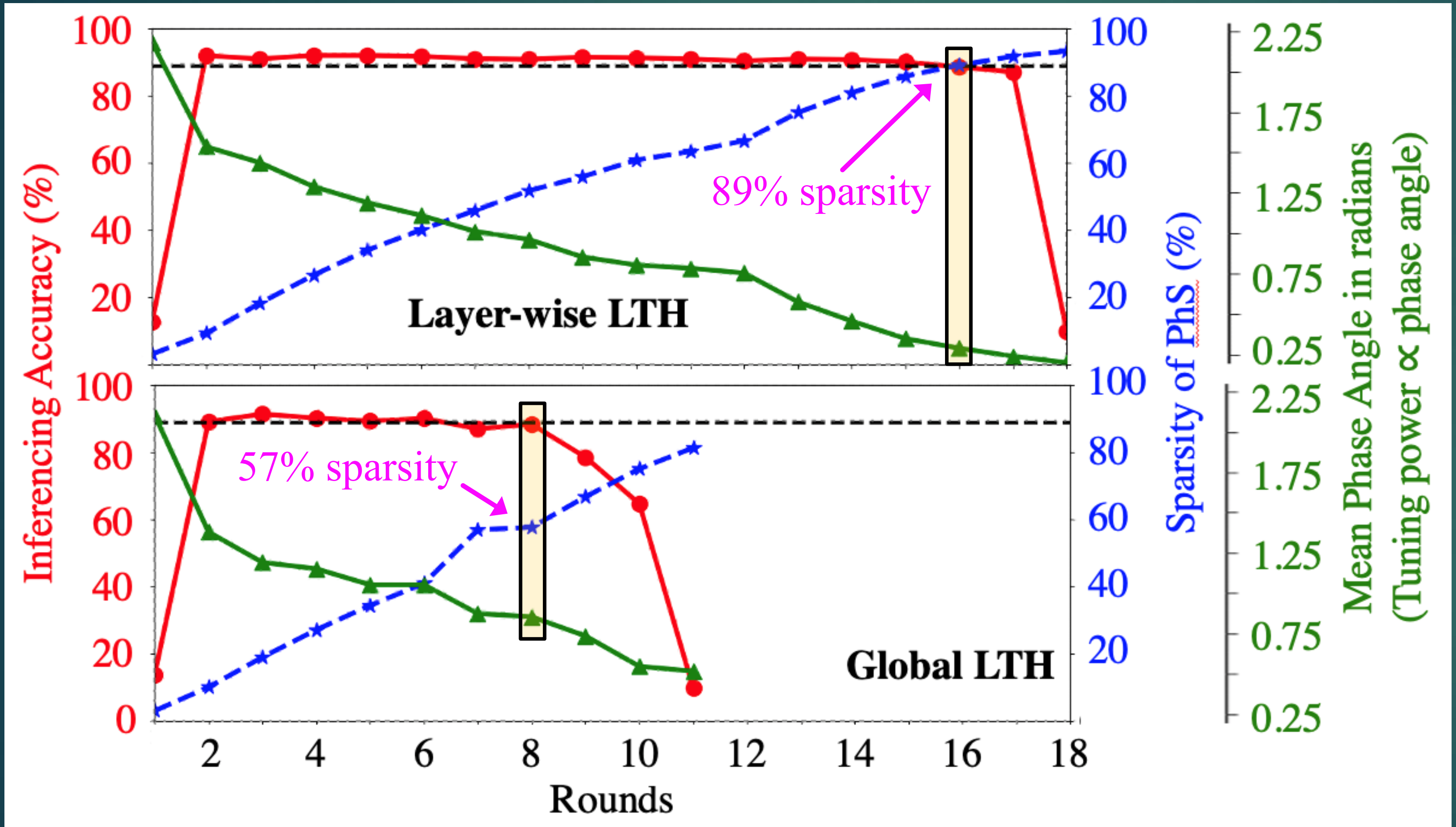


Acc. Loss (%)	Sparsity (%)	Power Savings (%)
0	74.86	46.05
1	98.57	97.62
5	99.45	98.23

Lottery Ticket Hypothesis-Based Pruning



Simulation Results – LTH-Based Pruning



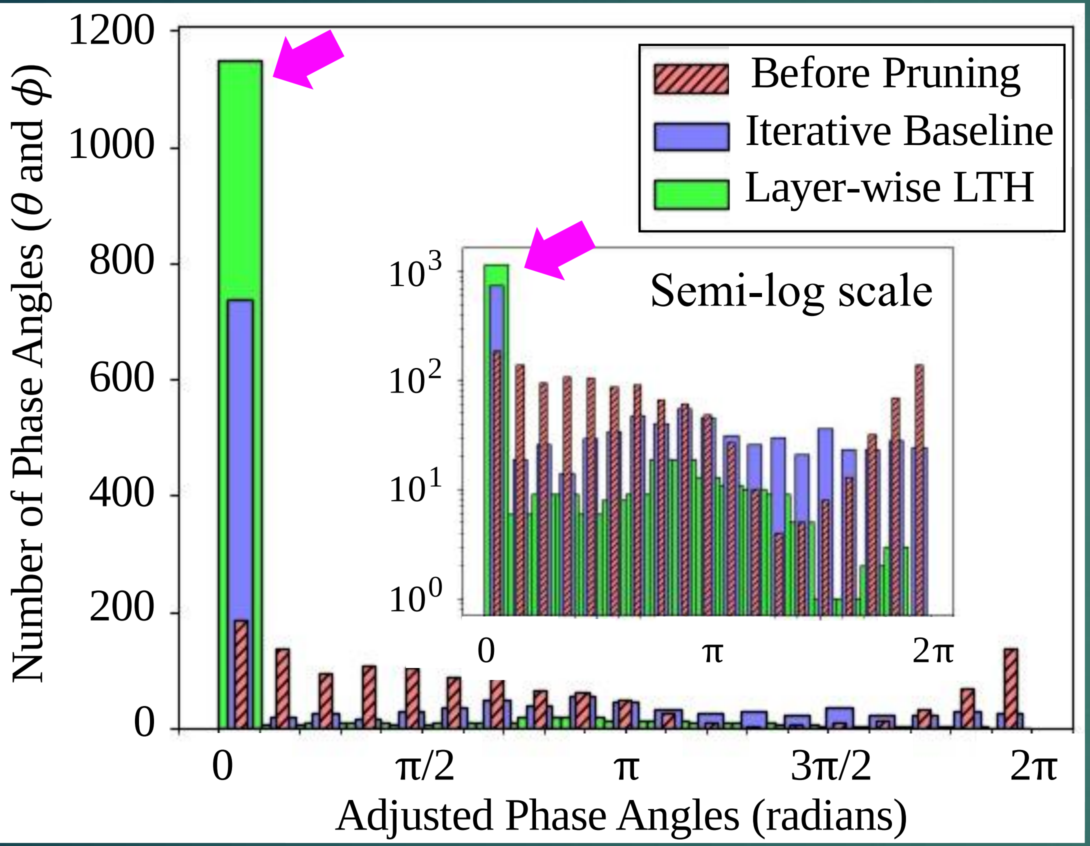
Layer-wise

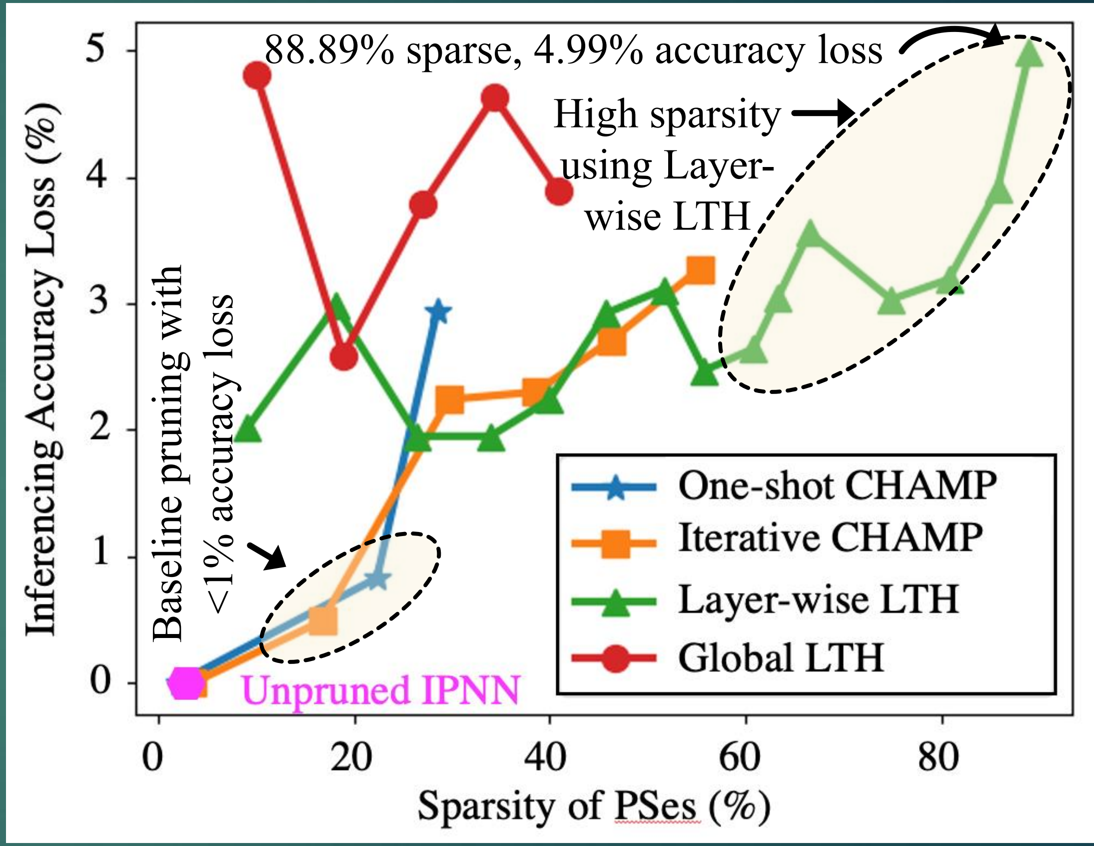
- Prune fraction of non-zero weights in each layer

Global

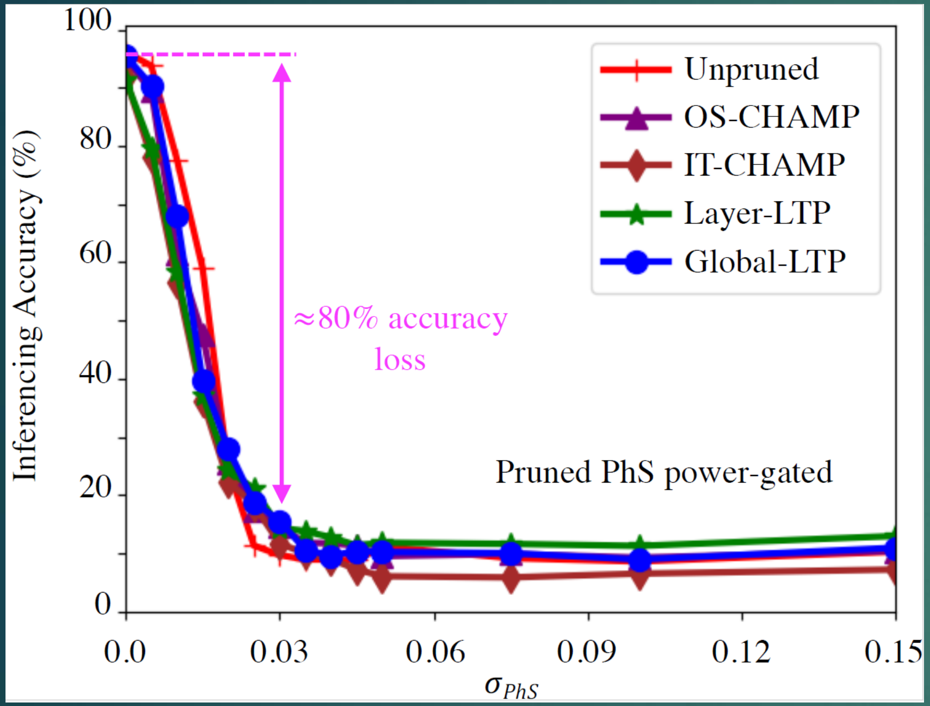
- Prune fraction of non-zero weights in the entire IPNN

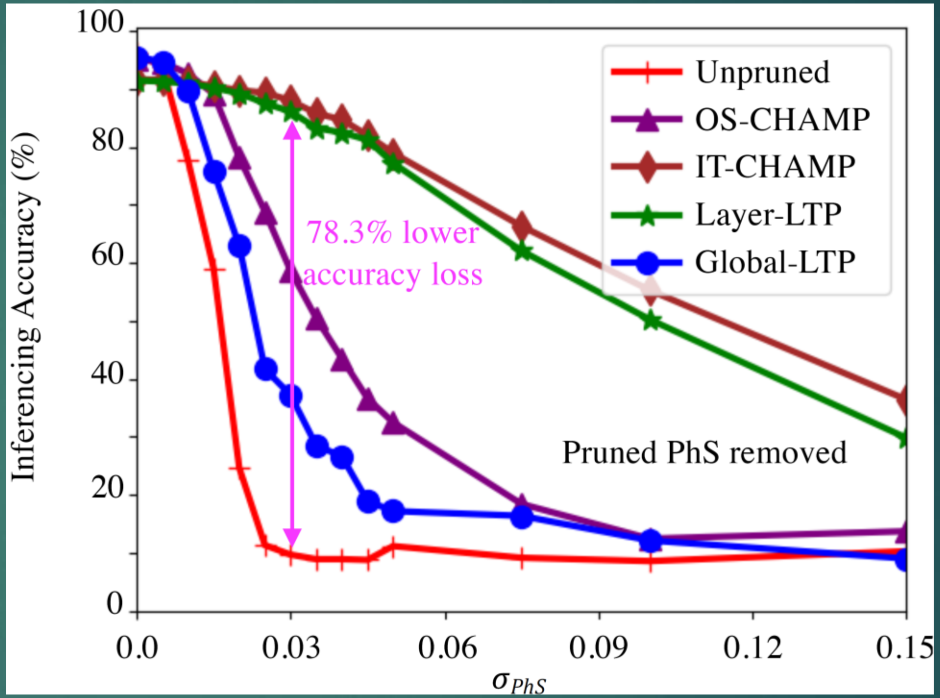
Simulation Results – LTH-Based Pruning





Simulation Results – LTH-Based Pruning

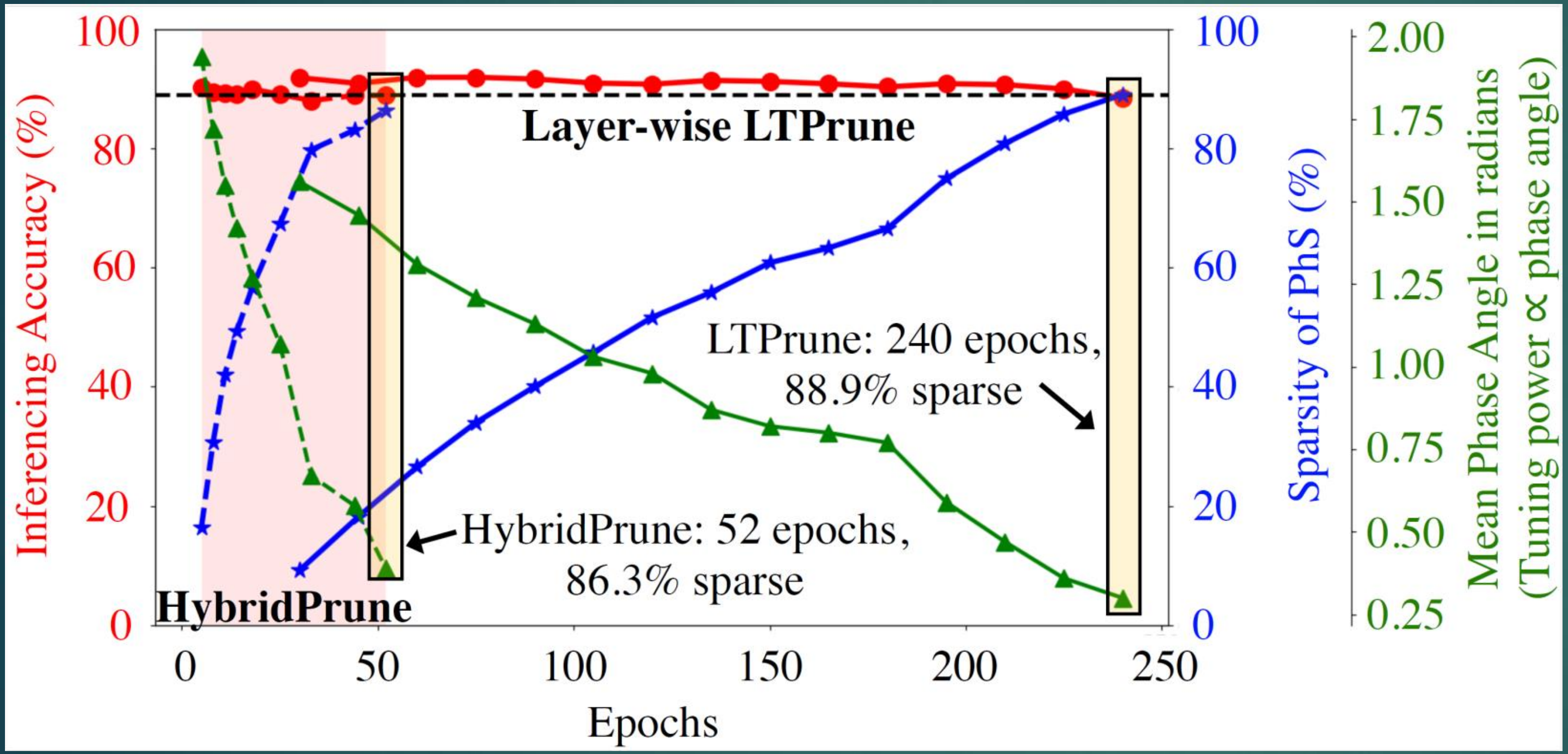




σ_{PhS}: Stdev. of Gaussian uncertainties in phase angles

Removing pruned phase shifters improves reliability

In-field Pruning using HybridPrune



Layer-wise LTPrune
- Slow but effective

Iterative CHAMP
- Quick but less sparse

HybridPrune
- Quick and effective

Key Takeaways

- Photonic NNs: ultra-fast low-energy matrix multiplication
- Correlated uncertainties in PhS and BeS are more critical
- MZIs in the initial IPNN layers are more critical
- Mitigative techniques should target PhS
- Pruning: reliability⬆, power⬇, footprint⬇
- Pruning for photonic NNs must be hardware-aware

Collaborations

- Prof. Mahdi Nikdast’s group and Prof. Sudeep Pasricha’s group from Colorado State University, Fort Collins

Prof. Krishnendu Chakrabarty

Prof. Mahdi Nikdast

Prof. Sudeep Pasricha

Amin Shafiee

Thank You!

sanmitrab@nvidia.com