Light-AI Interaction: Bridging Photonics and AI with Cross-Layer Hardware/Algorithm Co-Design

Jiaqi Gu[1,2], Hanqing Zhu[1], Chenghao Feng[1,3], Ray T. Chen[1], David Z. Pan[1]

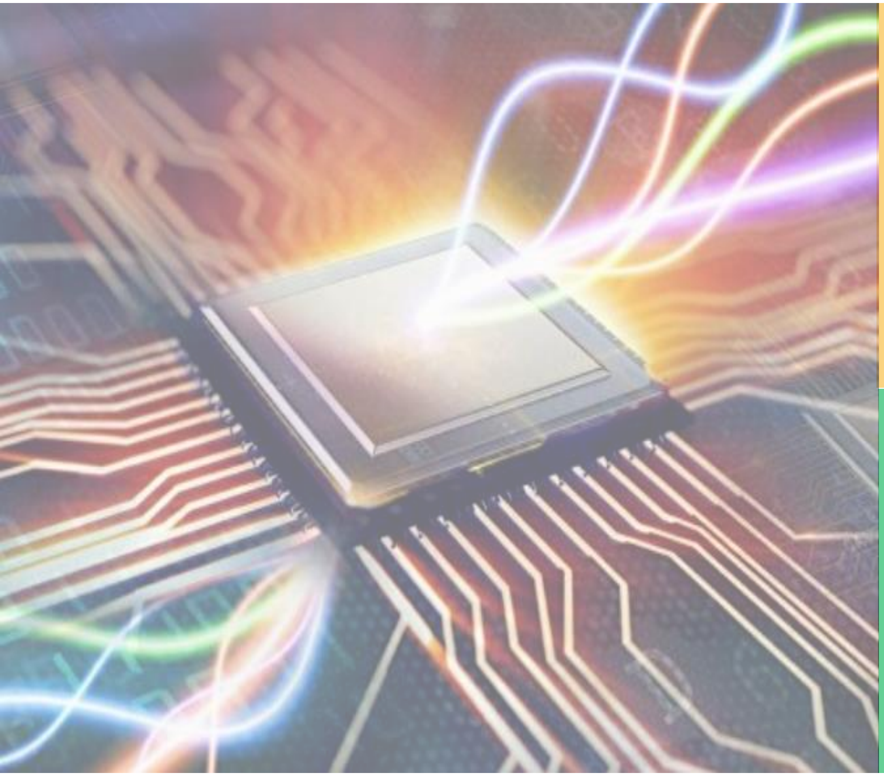
[1]The University of Texas at Austin, [2]Arizona State University, [3]Alpine Optoelectronics

jqgu@utexas.edu; http://jqgu.net

This work was supported in part by AFOSR MURI

June 15, 2023

What is This Talk About: Overview of Optical AI



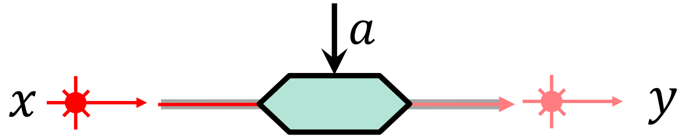
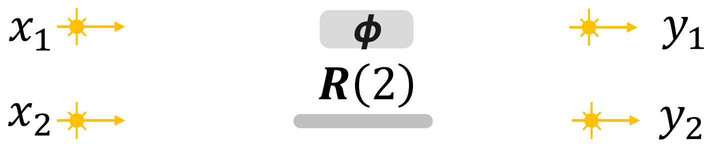
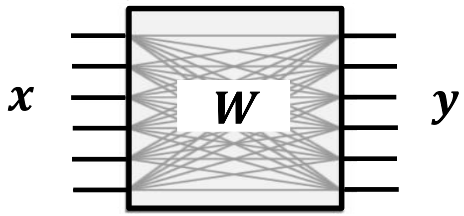
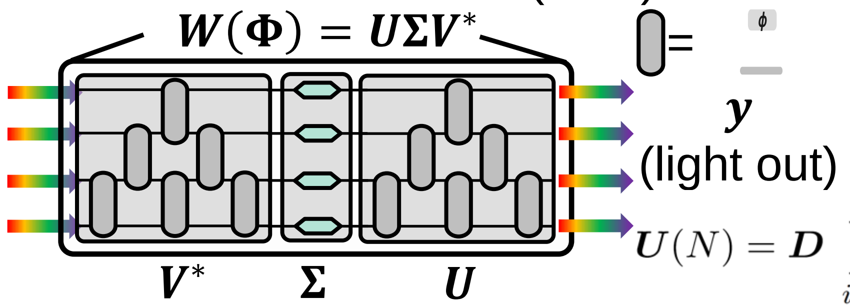
Design **Integrated Photonics Hardware**

for AI / ML Computing

Apply **AI / ML** for Photonic Hardware Design Automation

Photonic AI Computing Basics

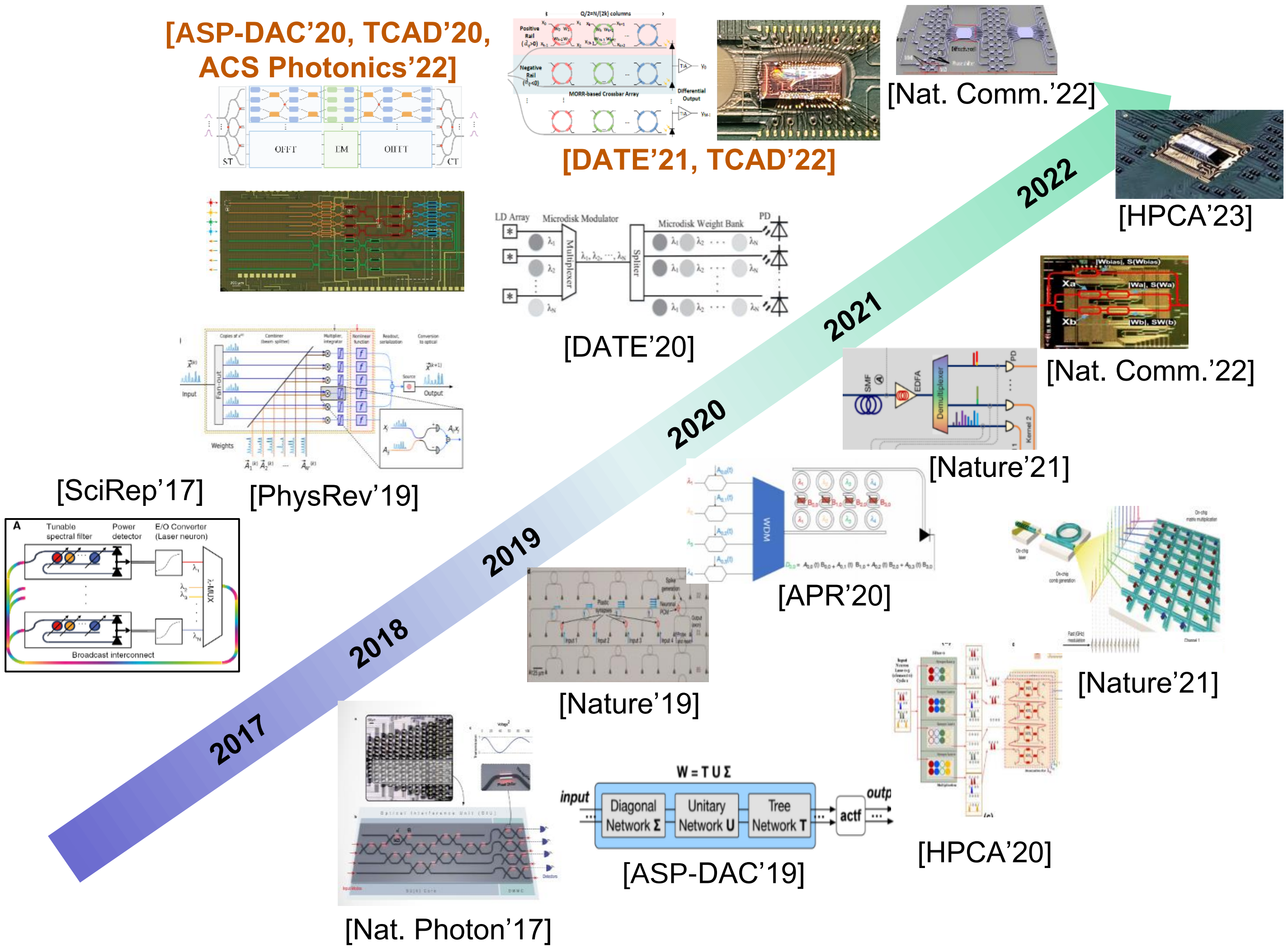
- Principle: modulation (encode), interference (MVM), photodetection (readout)
- Good at <u>ultra-fast</u> (10-100ps), <u>parallel</u> linear operations in the ***<u>analog</u>*** domain
- 10 TOPS/W (SoTA) → 1M TOPS/W (potential)

Computing Primitives	Photonic Implementation
Scalar Multiply $y = a \cdot x$	Light Modulation 
2×2 Unitary Matrix Multiply $\boldsymbol{y} = \boldsymbol{R}(2) \times \boldsymbol{x}$ $\boldsymbol{R}(2) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$	Mach-Zehnder Interferometer (MZI) 
Matrix-Vector Multiply (MVM) $\boldsymbol{y} = \boldsymbol{W} \times \boldsymbol{x}$  	Photonic Tensor Core (PTC)  $U(N) = \boldsymbol{D} \prod_{i=N}^{2} \prod_{j=1}^{i-1} \boldsymbol{R}_{ij}(\phi_{ij})$

One-shot computing at speed-of-light!

Photonic AI is Booming

Photonic Neural Network Trends in Academia

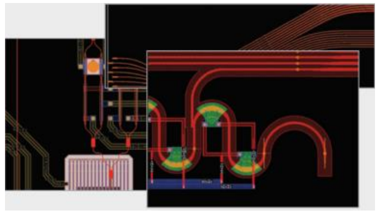


Foundry / EPDA Support in Industry

Photonic Computing Chip Designs

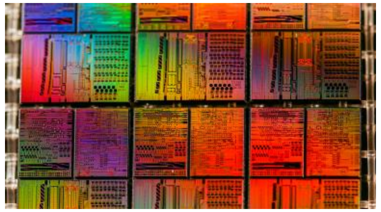
LIGHTMATTER
Lighton
XANADU
LIGHTELLIGENCE
LUMINOUS
OPTELLIGENCE
OPTIUS light powered computing
CogniFiber
SALIENCE LABS

Electronic-Photonic Design Automation Tools



Ansys LUMERICAL
SYNOPSYS PHOTONIC SOLUTIONS
cadence PHOTONICS

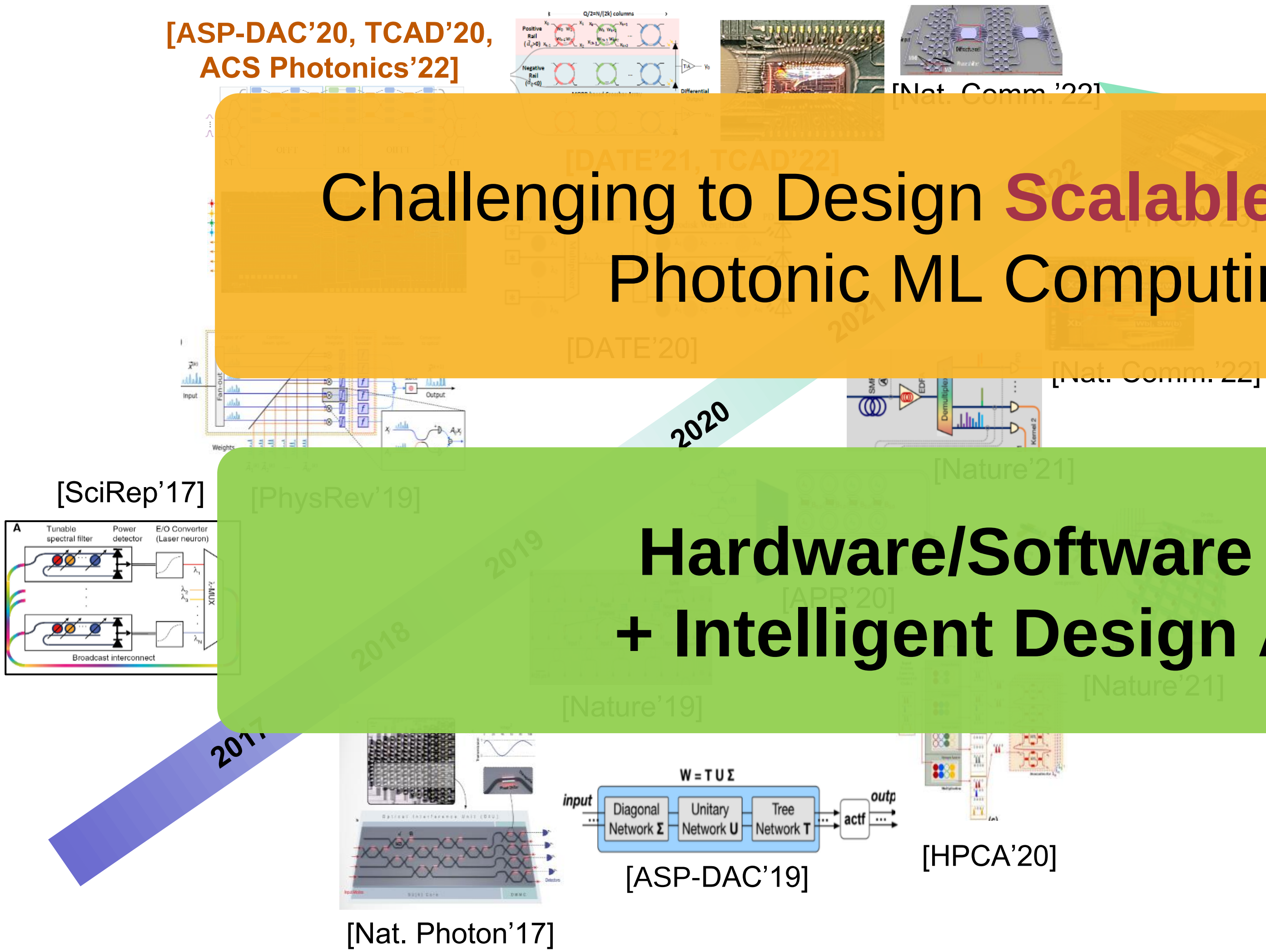
PDK / Tape-out / Packaging Support



amf ADVANCED MICRO FOUNDRY
AIM photonics
SiEPIC
GlobalFoundries
Tower Semiconductor

Photonic AI is Booming

Photonic Neural Network Trends in Academia

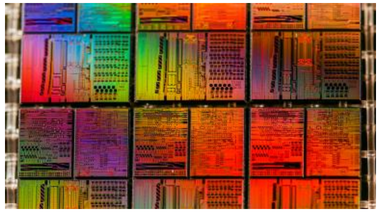


Foundry / EPDA Support in Industry

Photonic Computing Chip Designs

Electronic-Photonic Design Automation Tools

PDK / Tape-out / Packaging Support



Challenging to Design **Scalable, Robust & Adaptive** Photonic ML Computing Platforms

Hardware/Software Co-Design + Intelligent Design Automation

Virtuous Cycle: Photonics for AI ↔ AI for Photonics

Photonic AI Design Stack

- Photonic Computing Hardware Design
- Circuit-Model Co-Optimization
- Deployment & Application

Work & Contributions

Publications: >30 in CAD, ML, Photonics Communities

Area | Efficiency | Adaptability | Robustness

SqueezeLight, O²NN *[DATE'21] (**Tape-out**) [DATE'21]*

Optical RNN, Transformer *[CLEO'20] [MLSys SNAP'23]*

ADEPT *[DAC'22] (**Best-in-Track**)*

Mem-Efficient *[ICCV'21]*

Photonics+MTJ *[ICCAD'22]*

OSNN: Butterfly-style ONN *[ASP-DAC'20 **BPA**, TCAD'20, ACS Photonics'22] (**Tape-out**)*

Model Compression *[NeurIPS'22 MLSys, **Spotlight**]*

Robust ONN, Tutorial *[ICCAD'19,DATE'20] [TCAS II'21]*

NeurOLight *[NeurIPS'22, **Spotlight**]*

PCM-ONN *[ASP-DAC'22, TCAD'22]*

FLOPS, MixedTrain: Zeroth-order On-chip Train *[DAC'20, **BPC**] [NSF WS **BPA**] [AAAI'21]*

L²ight: Scalable On-chip Train *[NeurIPS'21]*

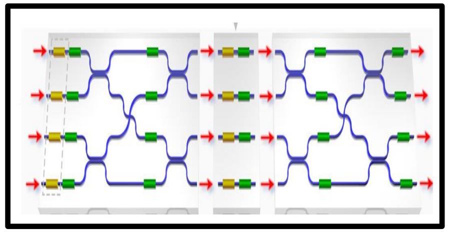
Open-Source Release

Torch ONN

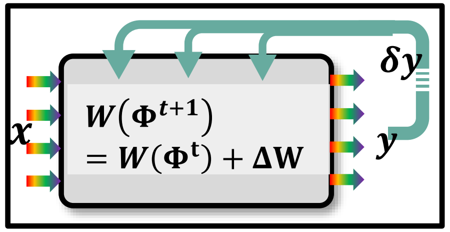
A PyTorch Library for Photonic Integrated Circuit Simulation and Photonic AI Computing

Fork 12 | Starred 205

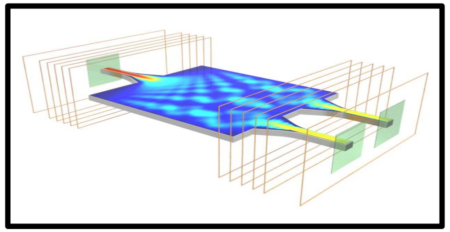
Outline



♦ Customized Optical Neural Network Design

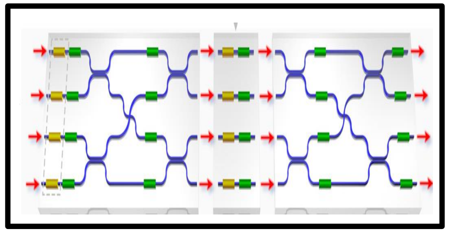


♦ ONN On-Chip Training Algorithms



♦ ML-Assisted Photonic Design Automation

Outline



- ♦ Customized Optical Neural Network Design

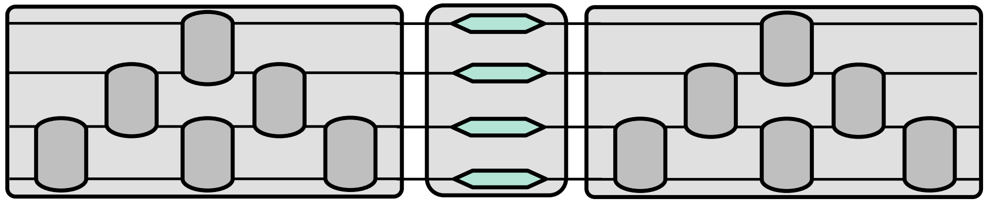
- ♦ ONN On-Chip Training Algorithms

- ♦ ML-Assisted Photonic Design Automation

From GEMM To Specialized Subspace Linear

- Overparameterized DNN → GEMM is not necessary → “circuit compression”
- **Large universal $U\Sigma V$** MZI array → Compact subspace $U'\Sigma V'$ butterfly mesh

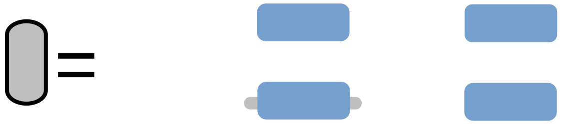
V^* Σ U



Instead of having general matrices…

$W_{ij} \times x_j$ $\quad O(2n^2)$

Large MZI array
$\mathcal{O}(n^2)$ MZIs



Break it to basic devices

50/50 coupler

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix}$$

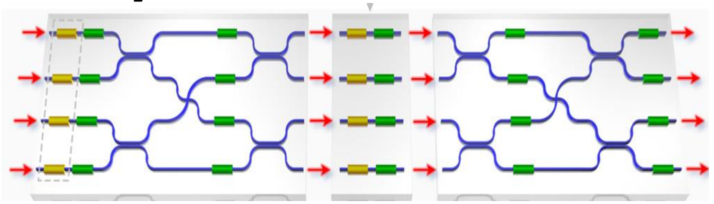
phase shifter

$e^{j\phi}$

Compact butterfly mesh
$\mathcal{O}(n \log n)$ basic devices

$\mathcal{F}^{-1}\left(\mathcal{F}(\boldsymbol{w_{ij}}) \odot \mathcal{F}(\boldsymbol{x_j})\right)$ $\quad O(n^2 \log_2 k / k)$

$V^{*\prime}$ Σ U'

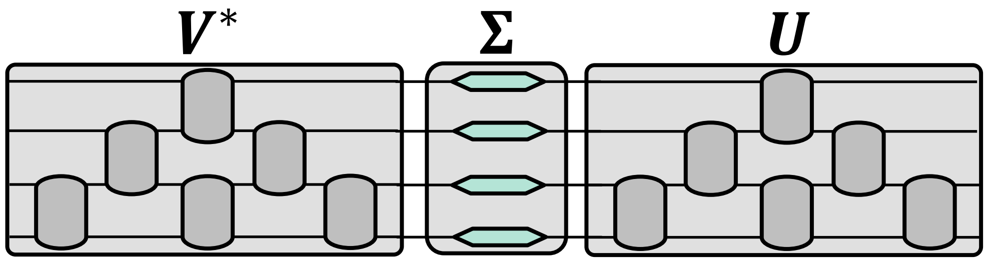


Subspace matrices

J. Gu, Z. Zhao, C. Feng, M. Liu, R.T. Chen, D.Z. Pan, “Towards Area-Efficient Optical Neural Networks: An FFT-based Architecture,” **ACM/IEEE ASP-DAC**, 2020. **Best Paper Award**

From GEMM To Specialized Subspace Linear

- Overparameterized DNN → GEMM is not necessary → “circuit compression”
- **Large universal $U\Sigma V$** MZI array → Compact subspace $U'\Sigma V'$ butterfly mesh

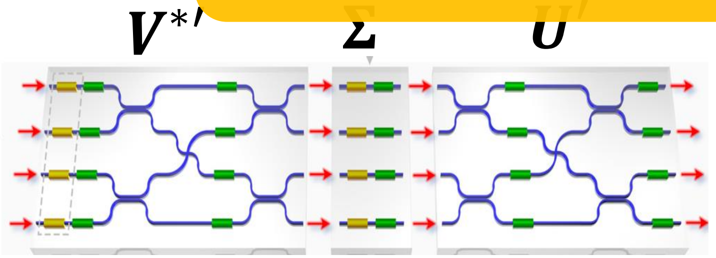


Instead of having general matrices…

$W_{ij} \times x_j$

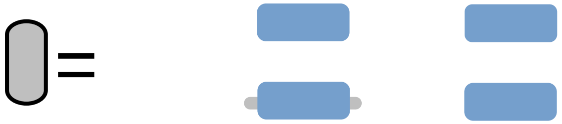
Trade *redundant matrix expressivity* for higher *hardware efficiency*

$\mathcal{F}^{-1}\left(\mathcal{F}(w_{ij}) \odot \mathcal{F}(x_j)\right)$ $O(n\log_2 k/k)$



Subspace matrices

Large MZI array
$O(n^2)$ MZIs



Break it to basic devices

Compact butterfly mesh

$O(n\log n)$ basic devices

50/50 coupler

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & j\\ j & 1\end{pmatrix}$$

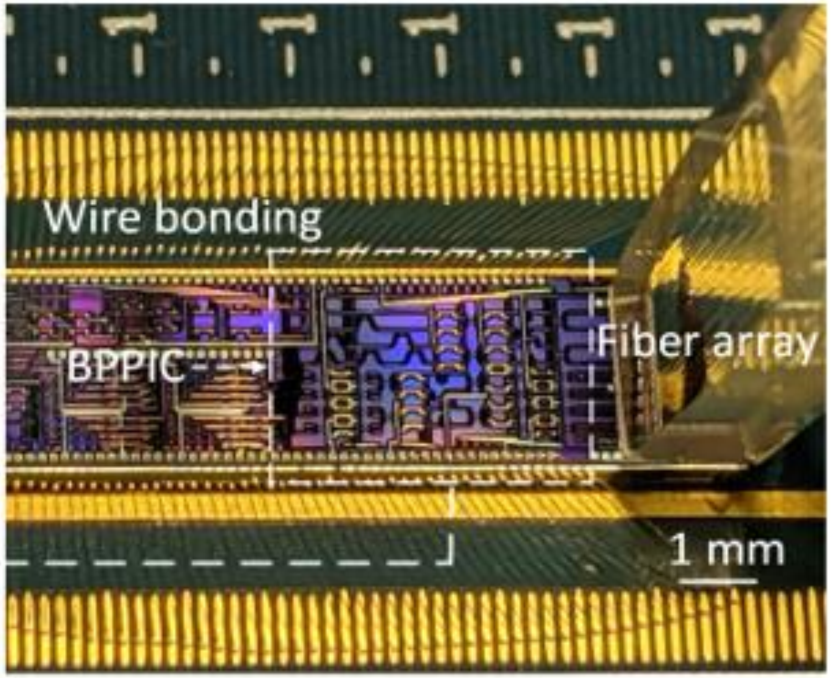
phase shifter

$e^{j\phi}$

J. Gu, Z. Zhao, C. Feng, M. Liu, R.T. Chen, D.Z. Pan, “Towards Area-Efficient Optical Neural Networks: An FFT-based Architecture,” **ACM/IEEE ASP-DAC**, 2020. **Best Paper Award**

Photonic Neural Chip Tapeout & Demonstration

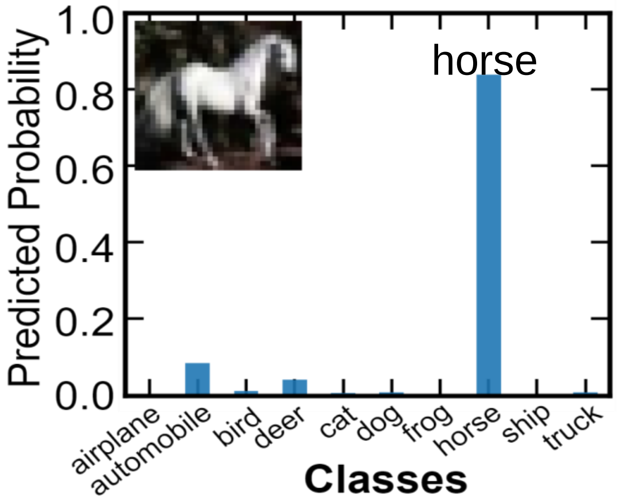
4×4 butterfly tensor core



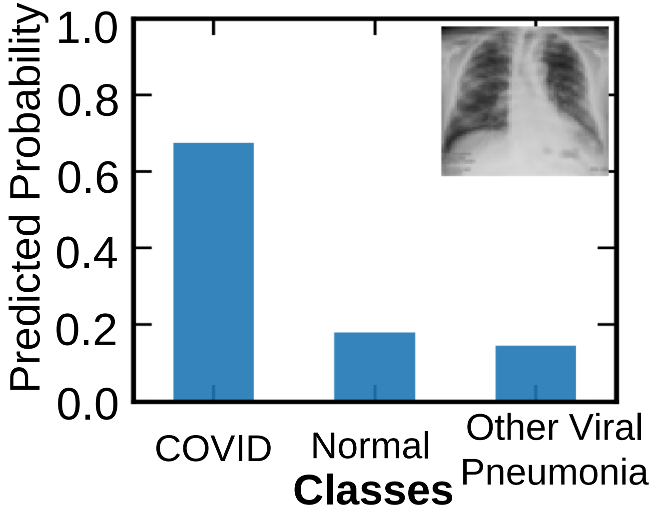
Peak compute density & energy efficiency

- **Our ONN** (225 TOPS/mm^2, 9.5 TOPS/W)
- NVIDIA A100 (0.76 TOPS/mm^2, 1.56 TOPS/W)
- Google TPUv4 (0.69 TOPS/mm^2, 1.62 TOPS/W)
- 40nm RRAM Accel (0.03 TOPS/mm^2, 2.20 TOPS/W) [Giordano+, VLSI'21]

>85% accuracy
ResNet-20 (0.27M #param) CIFAR-10
3-bit weight resolution
Fixed butterfly transform



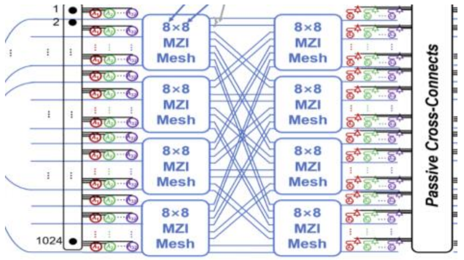
96.5% accuracy
VGG8 (4M #param) COVID Chest X-ray
3-bit weight resolution
Fixed butterfly transform



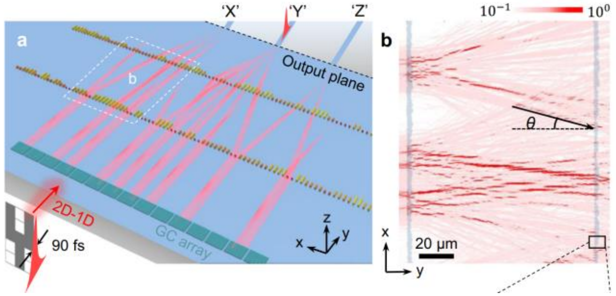
C. Feng*, J. Gu* (co-first), H. Zhu, Z. Ying, Z. Zhao, D.Z. Pan, R.T. Chen, "A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning", **ACS Photonics**, Nov. 30, 2022.

More Customized ONN Designs Beyond GEMM

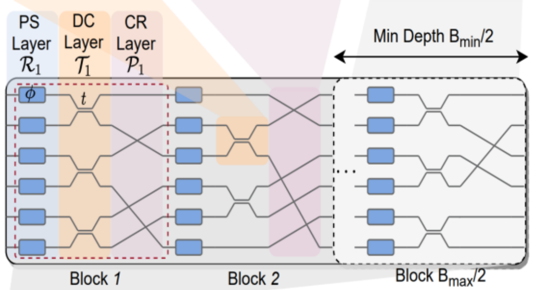
- **Specialized circuits** for hardware-efficient subspace linear op



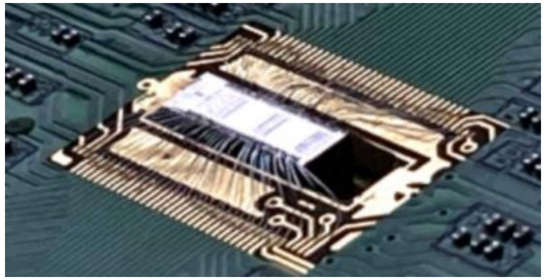
Tensorized MZI-ONN
[Xiao+, *APL Photonics* 2021]



Metalens-based diffractive ONN
[Wang+, *Nat. Commun* 2022]

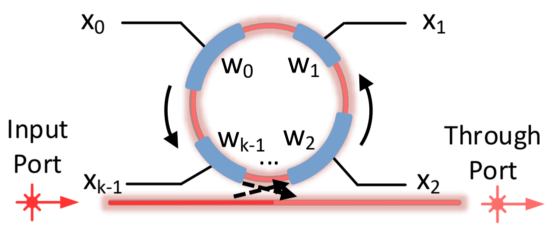


Auto-designed PIC topology
[Gu+, *DAC* 2022]

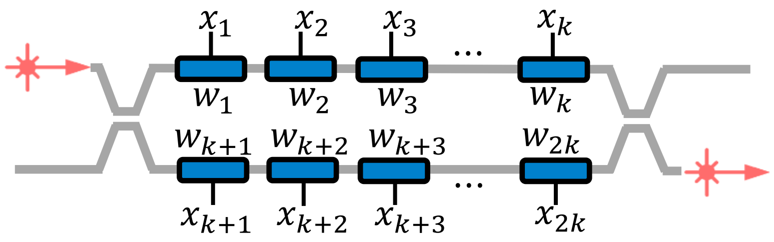


Fourier lens photonic Conv
[Li+, *HPCA* 2023]

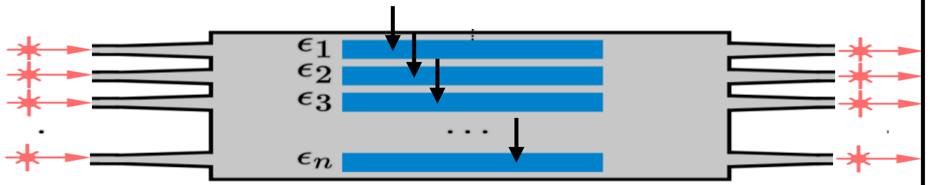
- **Customized devices** beyond 1 MAC/device → single-device vector/MVM unit



$y \propto f(\sum_{i=0}^{k-1} w_i x_i^2)$



$y \propto \cos(\sum_{i=1}^{2k} w_i x_i^2)$

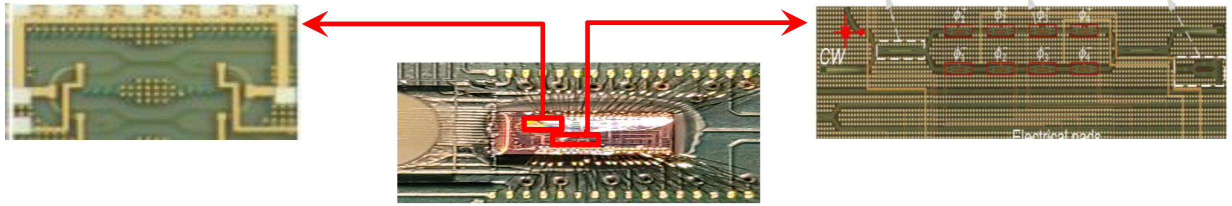


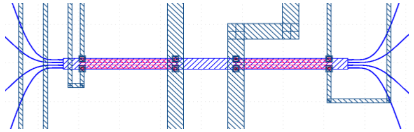
$f(\boldsymbol{x}, \boldsymbol{\epsilon}) = \boldsymbol{W}(\boldsymbol{\epsilon})\boldsymbol{x}$

MORR: [Gu+, DATE'21,TCAD'22]

MOMZI: [Feng+, Pho.West'23, arXiv'23]

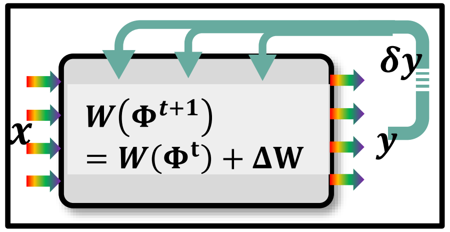
MOMMI: [Gu+, under sub., arXiv'23]





Outline

♦ Customized Optical Neural Network Design

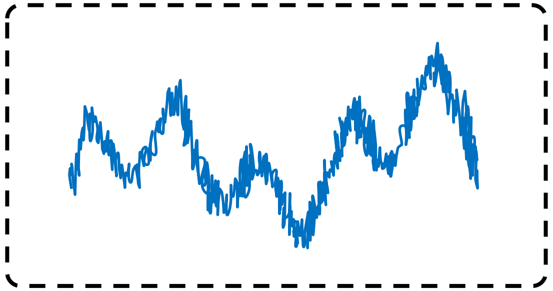


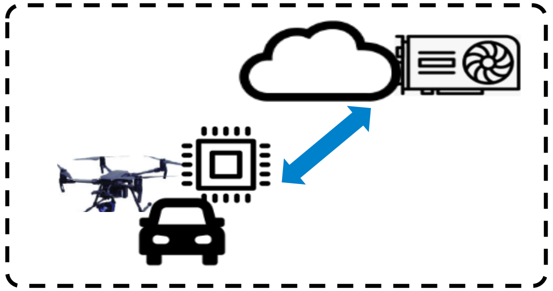
♦ ONN On-Chip Training Algorithms

♦ ML-Assisted Photonic Design Automation

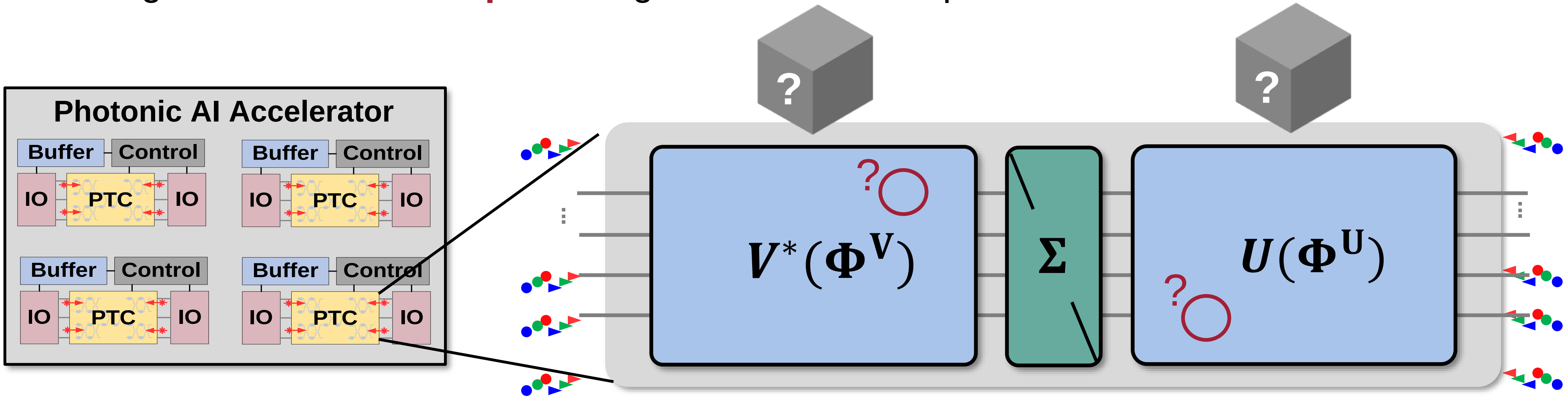
Inference → Training: Self-Learnable AI Engine

- Why on-chip training? *reliability, adaptability, efficiency, privacy…*



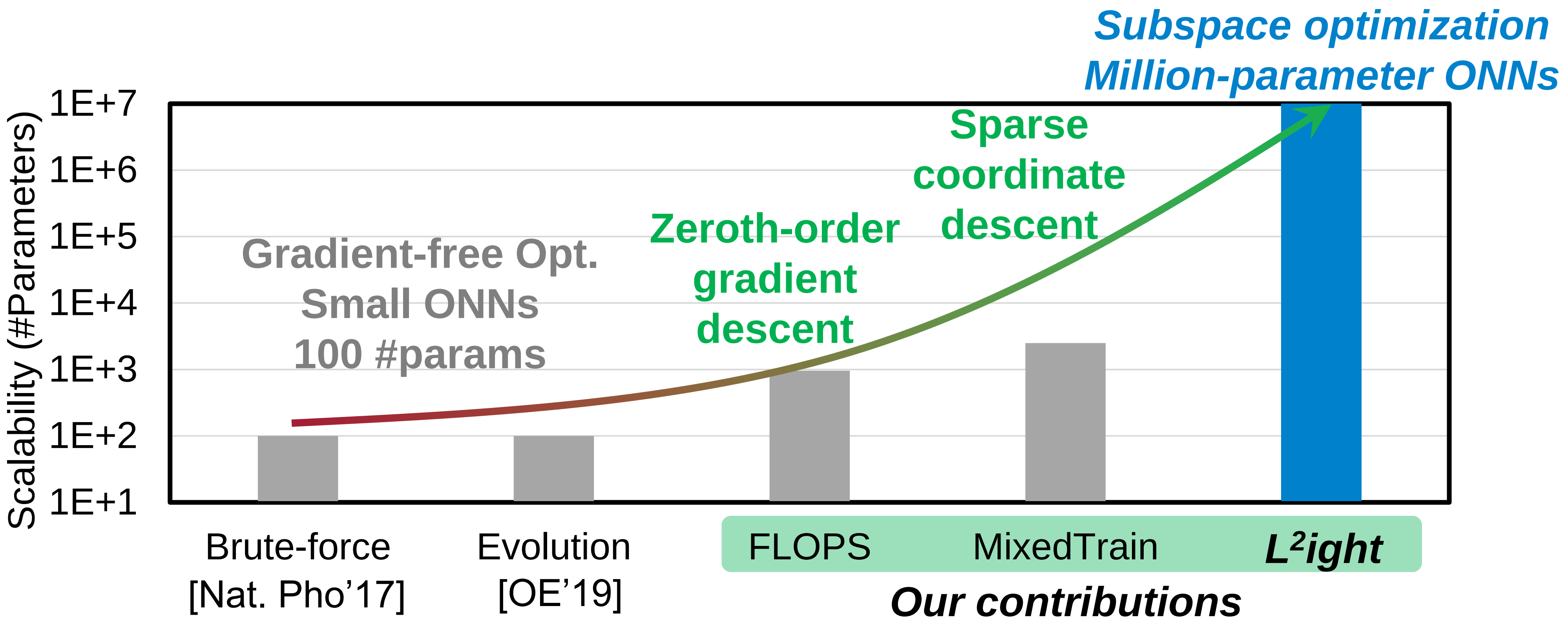


- Challenges
  - No access to **intermediate states or full gradients** (U/V are blackbox)
  - **Noisy** circuits (randomness)
  - Algorithm must be **simple** enough to be run on chip



Efficient On-Chip Training Protocols

- >10,000× trainability↑+30× efficiency↑: Customize algorithm for the hardware
- Utilize optics reciprocity to calculate *subspace* 1st-order gradients with *sparsity*



J. Gu, Z. Zhao, *et al.*, **FLOPS**, ACM/IEEE Design Automation Conference (**DAC**), 2020 (**Best Paper Finalist**) (**Best Poster Award**)

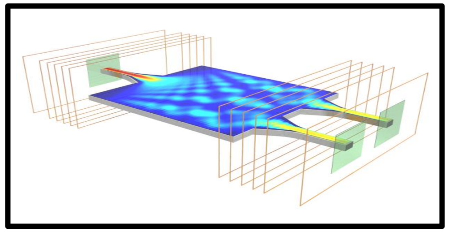
J. Gu, C. Feng, *et al.*, **Mixed-Train**, Association for the Advancement of Artificial Intelligence (**AAAI**), 2021

J. Gu, H. Zhu, *et al.*, **L2ight**, Conference on Neural Information Processing Systems (**NeurIPS**), 2021

Outline

♦ Customized Optical Neural Network Design

♦ ONN On-Chip Training Algorithms

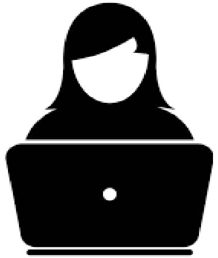


♦ ML-Assisted Photonic Design Automation

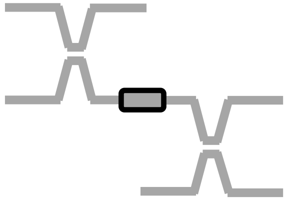
Photonics for AI

Light-AI
Virtuous Cycle

AI for Photonics

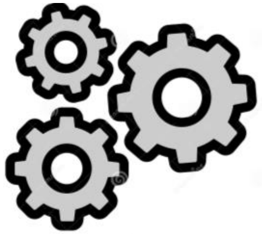


Manual Design

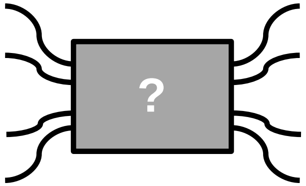


Standard Devices

AI-Enabled



<u>Automated</u> Photonic IC Design

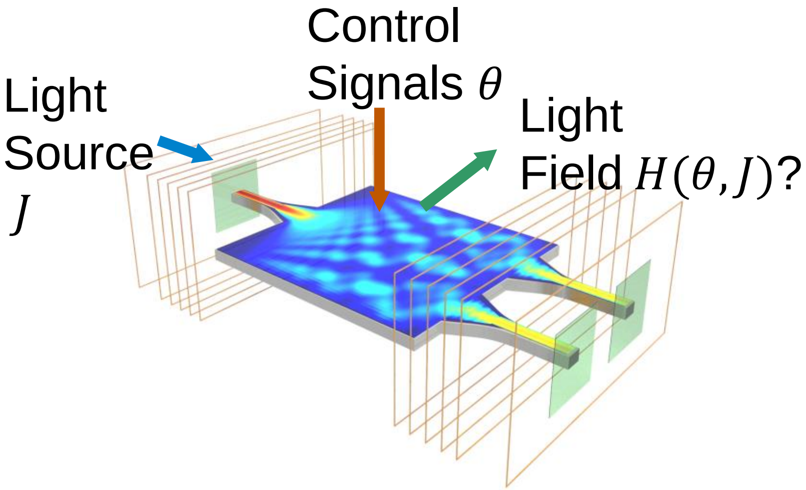


<u>Customized</u> Photonic Structure

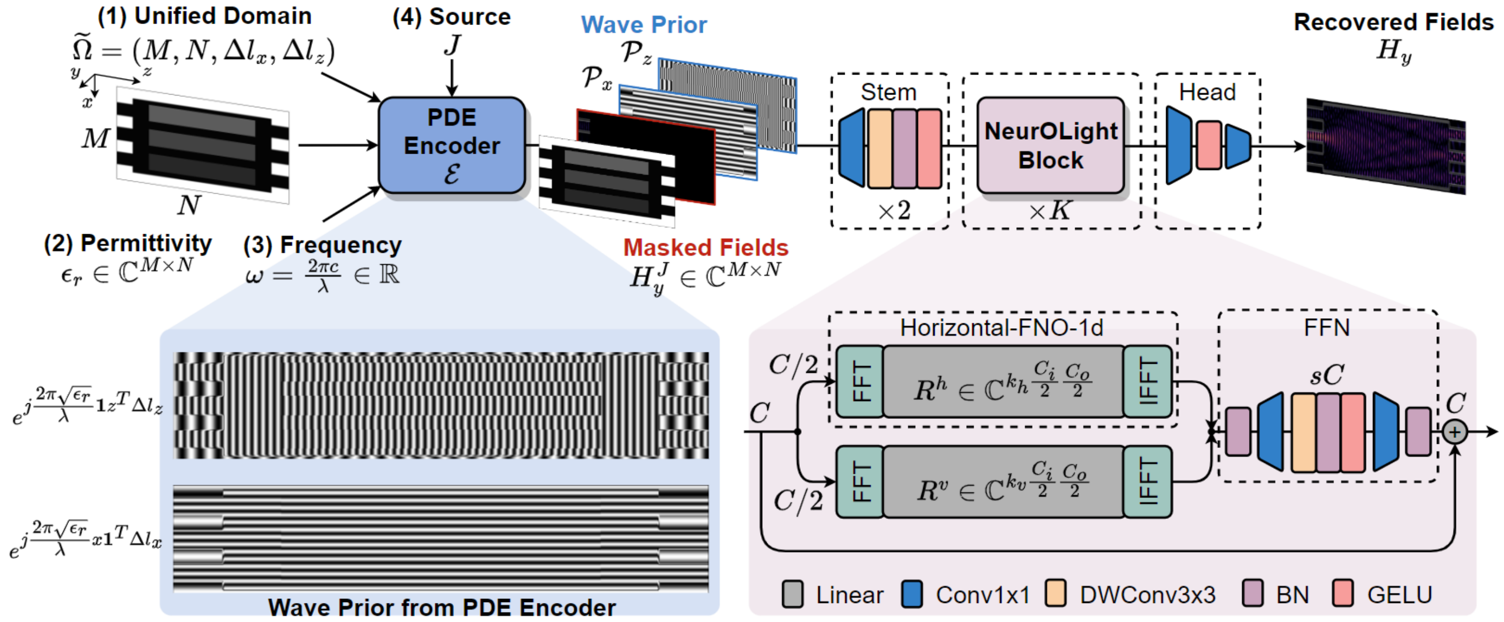
AI for Optical Simulation [NeurOLight, Gu+, NeurIPS'22]

- Avoid slow simulation in the loop → ML-enabled fast Maxwell equation solving

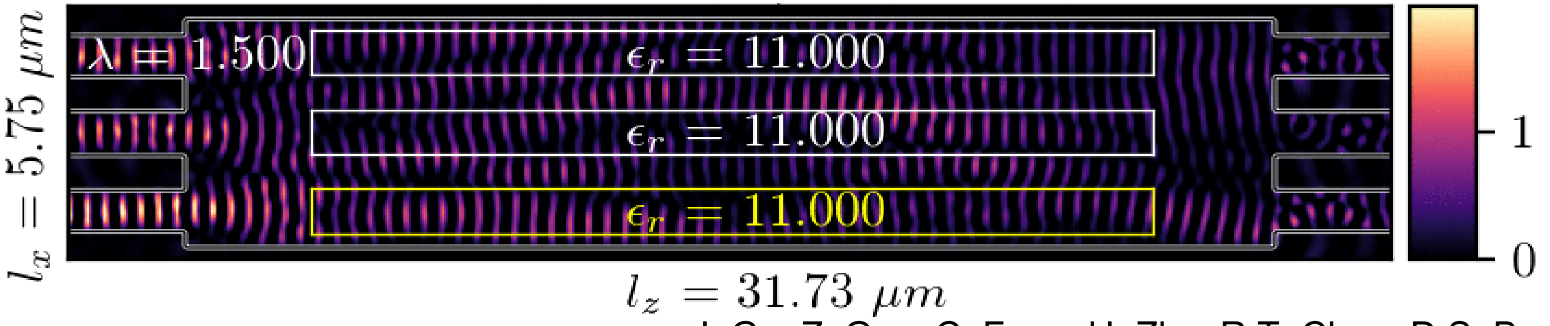
Obtain customized device behavior



PDE encoding + *efficient neural operator* + *physics-augmented training*



One-shot batched prediction on parametric Maxwell Eq. solutions

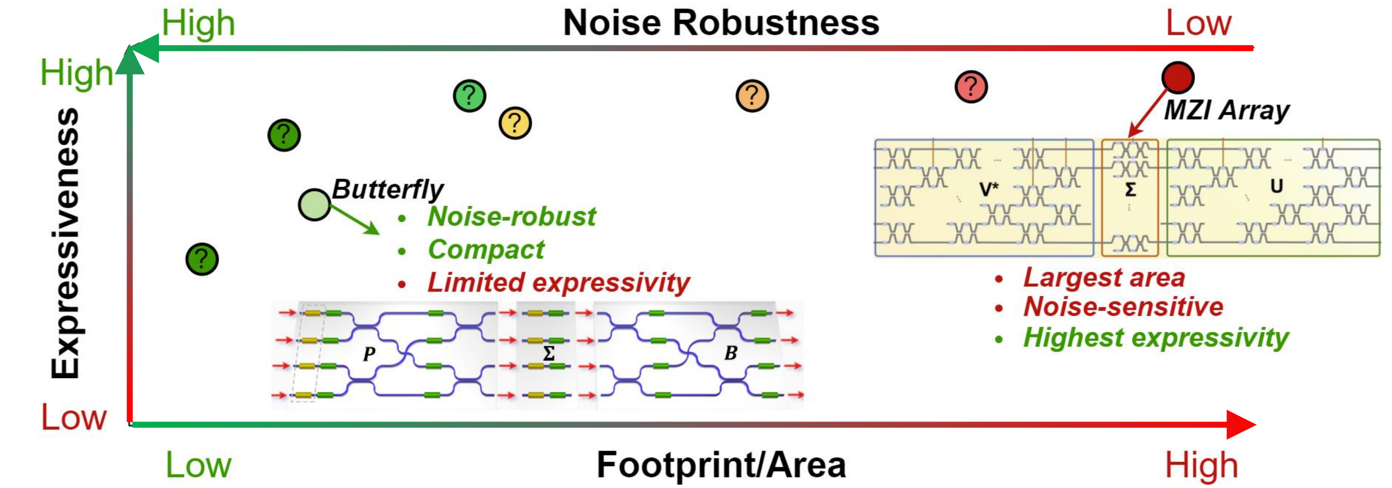


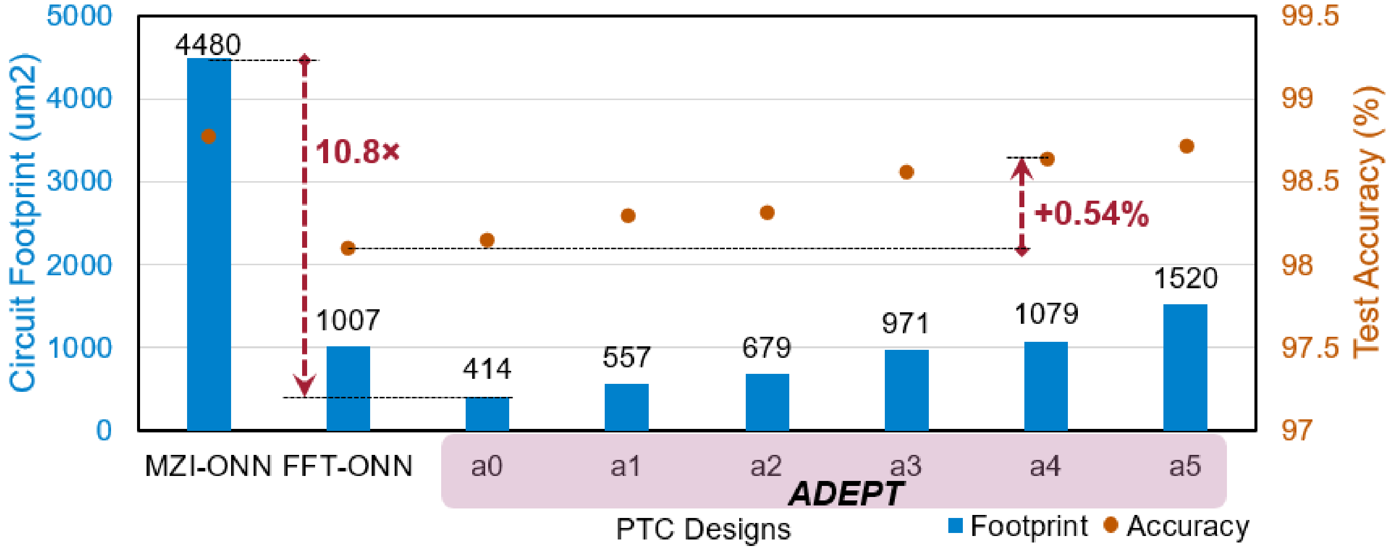
>200× speedup:
Fast inference (<10 ms)
120 FPS

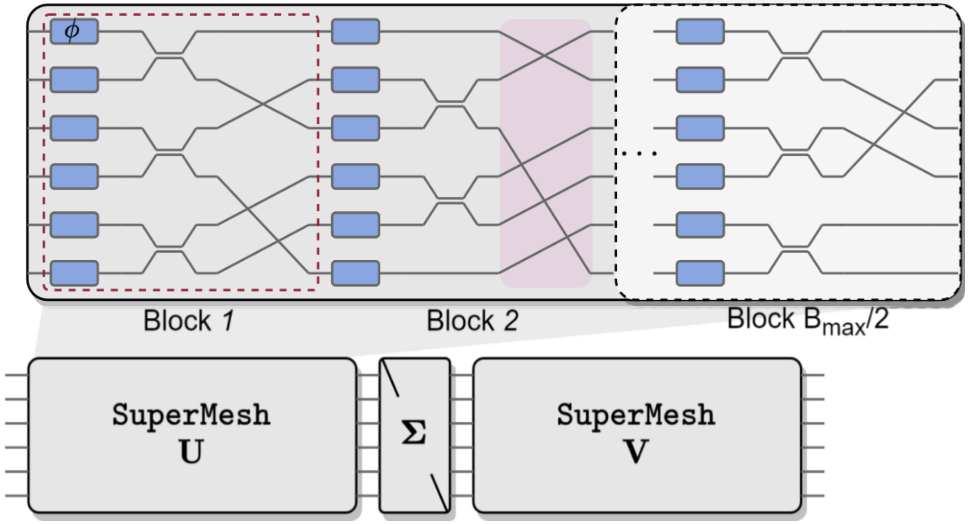
J. Gu, Z. Gao, C. Feng, H. Zhu, R.T. Chen, D.S. Boning, D.Z. Pan, "NeurOLight: A Physics-Agnostic Neural Operator Enabling Parametric Photonic Device Simulation," **NeurIPS** 2022. (**Spotlight**)

Auto-Design for Photonic Circuits [ADEPT, Gu+, DAC'22]

- Inefficient manual/heuristic design → Automated circuit topology search







Discrete → Continuous & Differentiable

Auto adapt to PDK and chip constraint

2-30× smaller
More noise-robust

J. Gu, H. Zhu, C. Feng, Z. Jiang, M. Liu, S. Zhang, R.T. Chen, D.Z. Pan, "ADEPT: Automatic Differentiable DEsign of Photonic Tensor Cores," **DAC**, 2022 (**Best-in-Track**)

The Future of Photonics↔AI is Bright

TorchONN

- HW/SW co-design for optical AI infer. /train + ML for optics
- Future: mem/arch, system integration, advanced app.

ONN Design Stack

Optical Neural Architecture Design

↓

ONN Model Optimization

↓

Deployment & On-Chip Training

Scalability ↑
Robustness ↑
Efficiency ↑
Trainability ↑

TorchONN

GPU-Backend

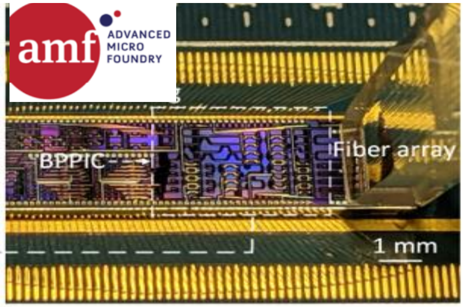
Co-Design Methodology

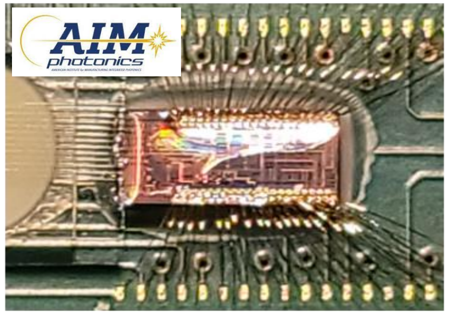
SOFTWARE

- ONN Architecture Search
- Automated ONN Core Design
- Area-Efficient ONN Arch Design
- Robust ONN Training
- ONN Pruning/Quantization
- Device Exploration
- Model Mapping
- On-Chip Learning

HARDWARE

JeremieMelo/pytorch-onn

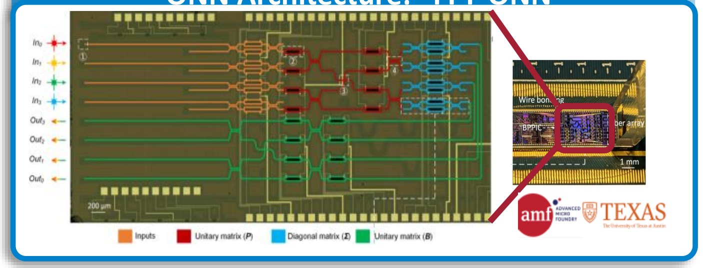




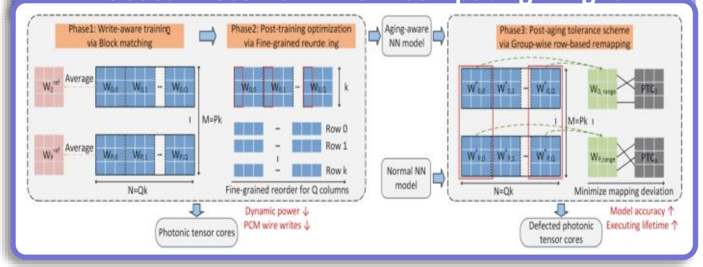
Photonics for AI
AI for Photonics

Hands-on Tutorial on TorchONN @ Design Automation Conference (DAC) July'23, Moscone Center

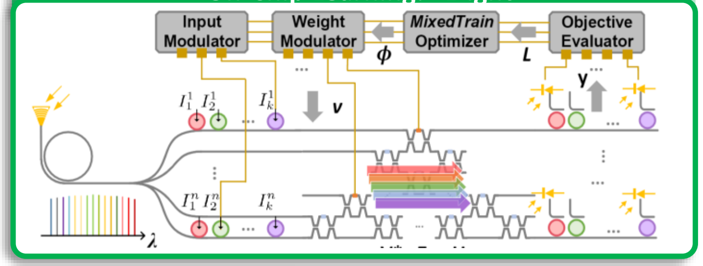
ONN Architecture: FFT-ONN



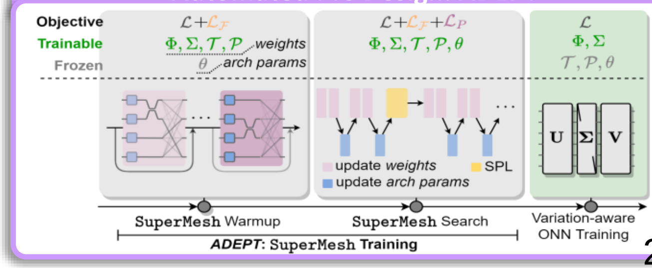
Robust Photonic In-Mem Computing: ELight



On-Chip Learning: L²ight



Automated PIC Design: ADEPT



Acknowledgment

Contributors: Jiaqi Gu, Chenghao Feng, Hanqing Zhu, Zheng Zhao, Zhoufeng Ying, Ray T. Chen, David Z. Pan

Funding Agency: AFOSR MURI

Thank You

Q & A