

SCATTER: Algorithm-Circuit Co-Sparse Photonic Accelerator with Thermal-Tolerant, Power-Efficient In- situ Light Redistribution

Dennis Yin¹, Nicholas Gangi², Meng Zhang²,
Jiaqi Gu¹, Jeff Zhang¹, Rena Huang²

¹Arizona State University, ²Rensselaer Polytechnic Institute

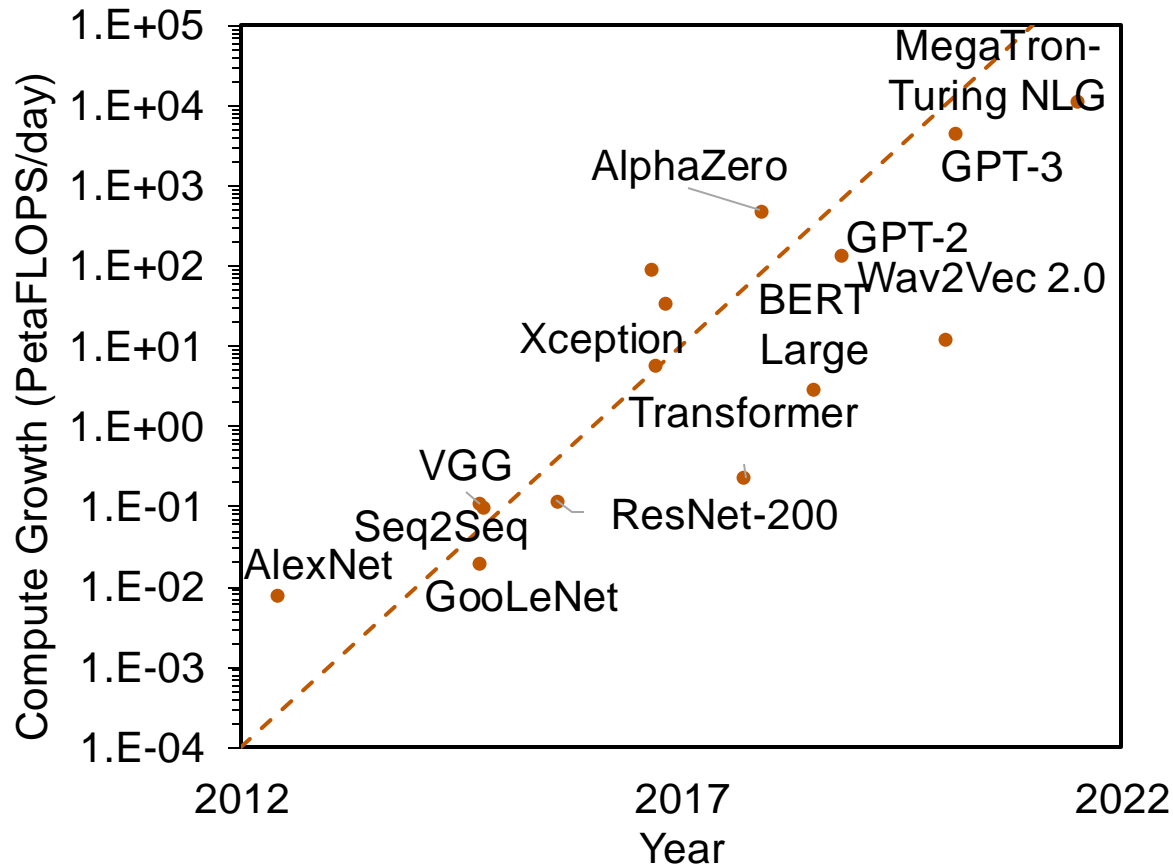
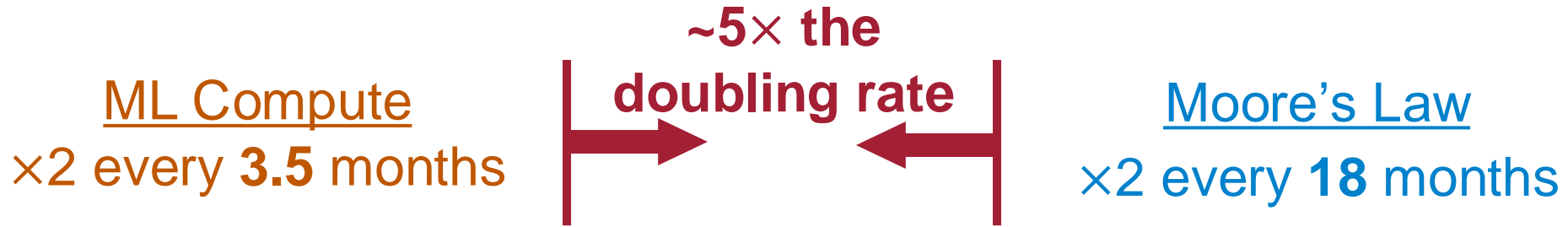
¹School of Electrical, Computer and Energy Engineering

zhangyin@asu.edu

jiaqigu@asu.edu | scopex-asu.github.io



Hardware Limits in Machine Learning



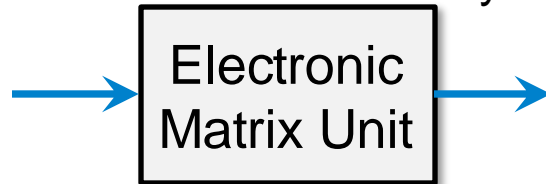
Need specialized emerging hardware and co-design methodology for speed and efficiency breakthrough

Source: <https://openai.com/blog/ai-and-compute/>
Source: <https://spectrum.ieee.org/nvidias-next-gpu-shows-that-transformers-are-transforming-ai>

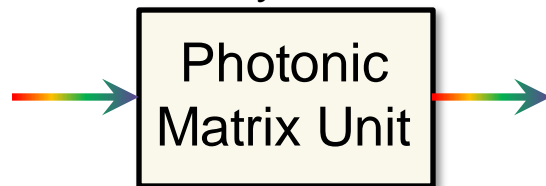
Electrical Computing vs Photonic Computing

High speed

Delay $100\text{ ns} \sim 1\ \mu\text{s}$
A few hundred clock cycles

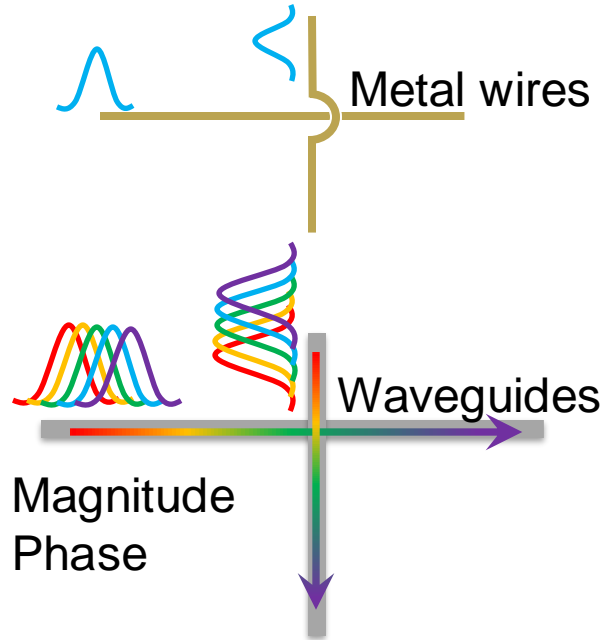


Delay $\ll 1\text{ ns}$



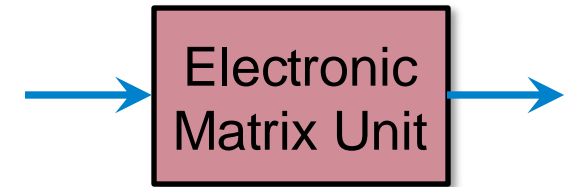
Computing as light propagate

Massive parallelism



Light propagate in parallel

High energy efficiency

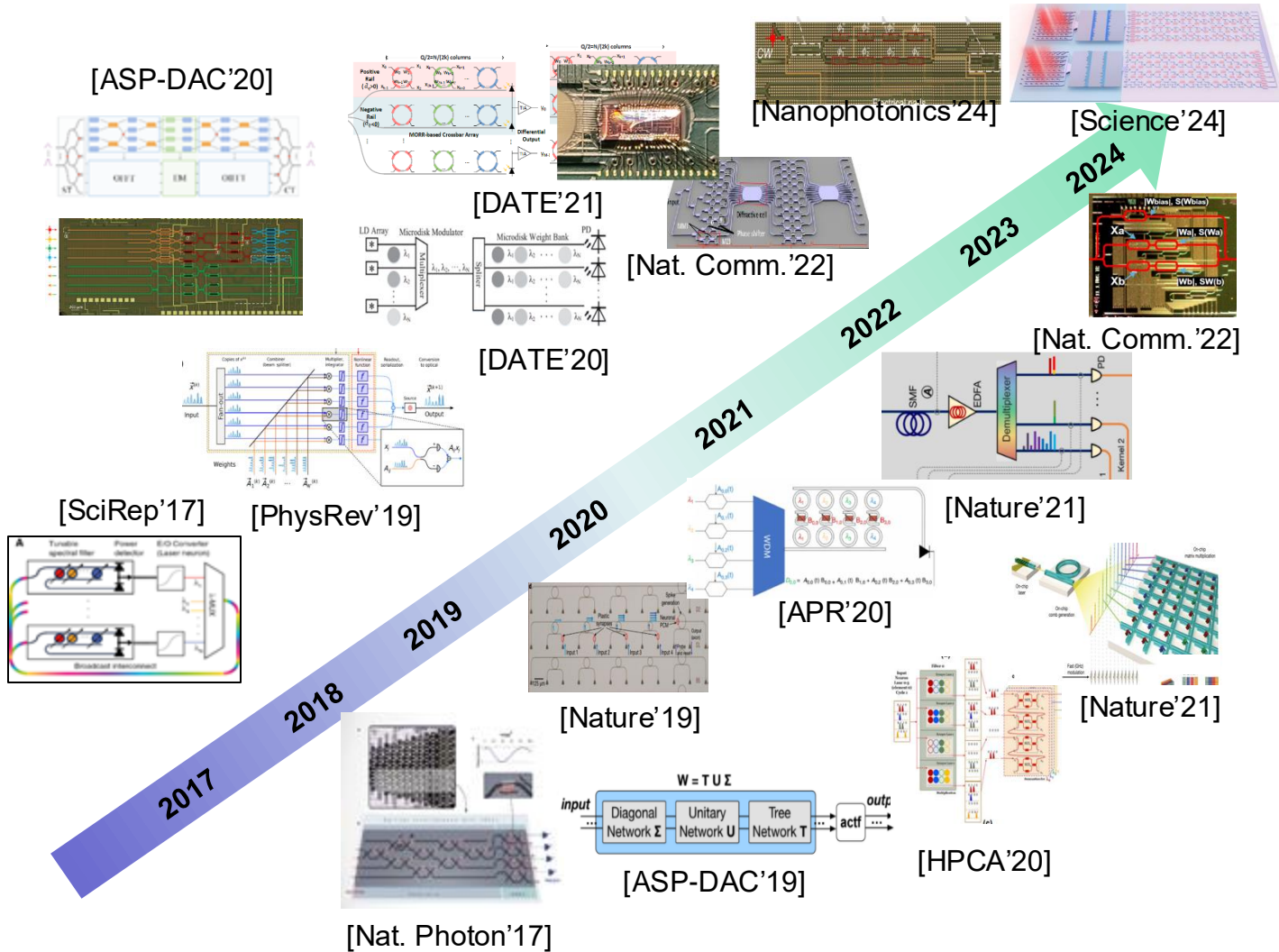


Passive circuits consumes near zero static power

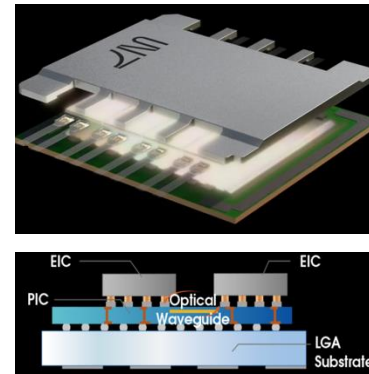
Photonic AI System is Booming

Photonic AI Trends in Academia

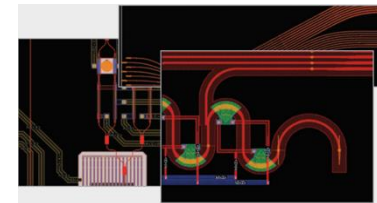
Foundry / EPDA Support in Industry



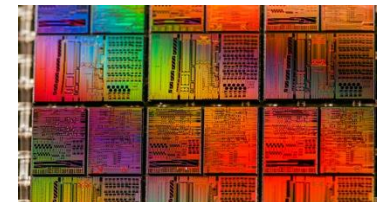
Photonic Computing Chip + Optical Interconnects



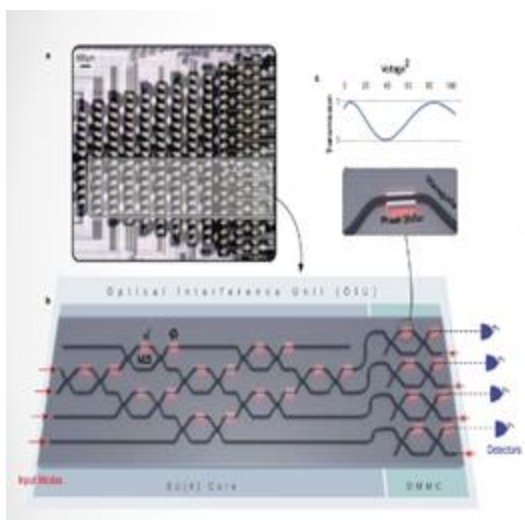
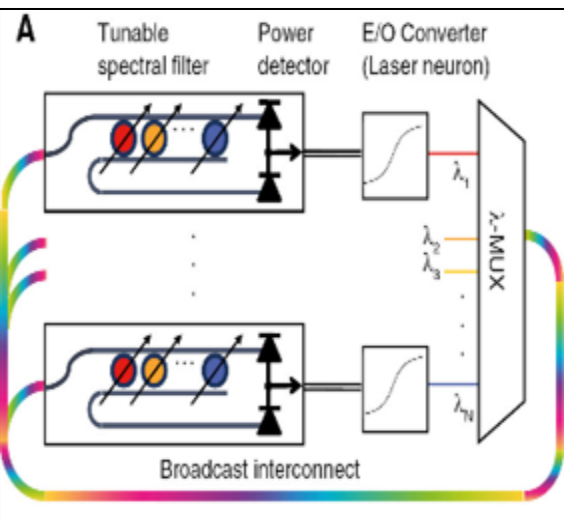
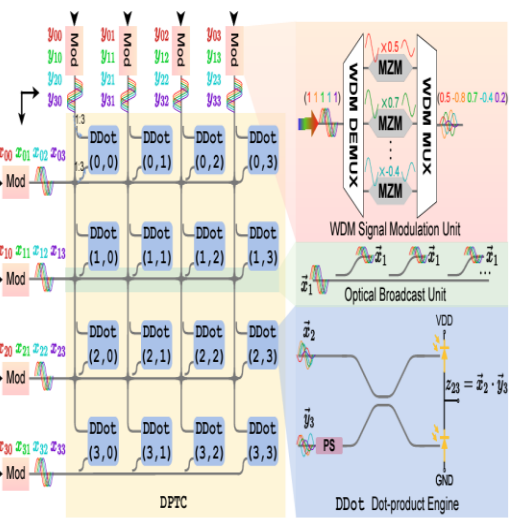
Electronic-Photonic Design Automation Tools



PDK / Tape-out / HI / E-O Co-Packaging Support



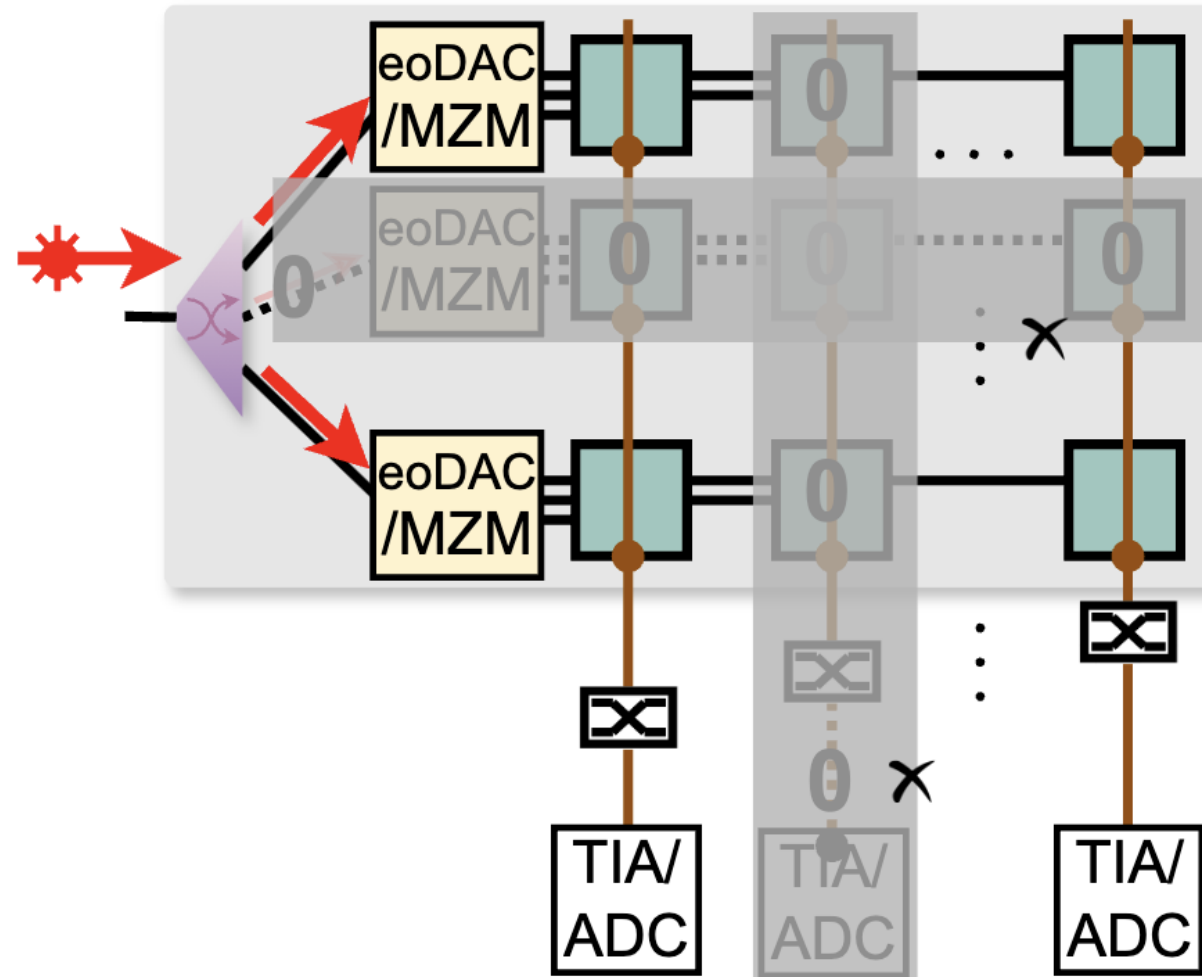
SCATTER Advantages Over Other Types of PTCs

		MZI Mesh [Shen+, NatPhot'17]	MRR Weight Bank [Tait+, SciRep'17]	Coupler Xbar [Zhu+, HPCA'24]	Common Challenges
					<ul style="list-style-type: none"> • Power-consuming E-O conversion • Large chip area cost • Analog device/circuit noise
Phase Sensitivity		Cumulative phase errors	Incoherent PTC	Hard to maintain $\frac{\pi}{2}$ phase change	
Thermal Sensitivity		Upper/lower arms cancel out	$< 1K$	$< 10K$	
Operand Range	A	\mathbb{R}	\mathbb{R}^+	\mathbb{R}	
	B	\mathbb{R} (SVD)	\mathbb{R}	\mathbb{R}	

Proposed Reconfigurable Sparse Photonic Accel.

- Lacks universality
- Phase/Thermal Sensitive
- Need large device spacing to reduce crosstalk
- Analog device/circuit noise
- Large on-chip area cost
- Power-consuming E-O conversion

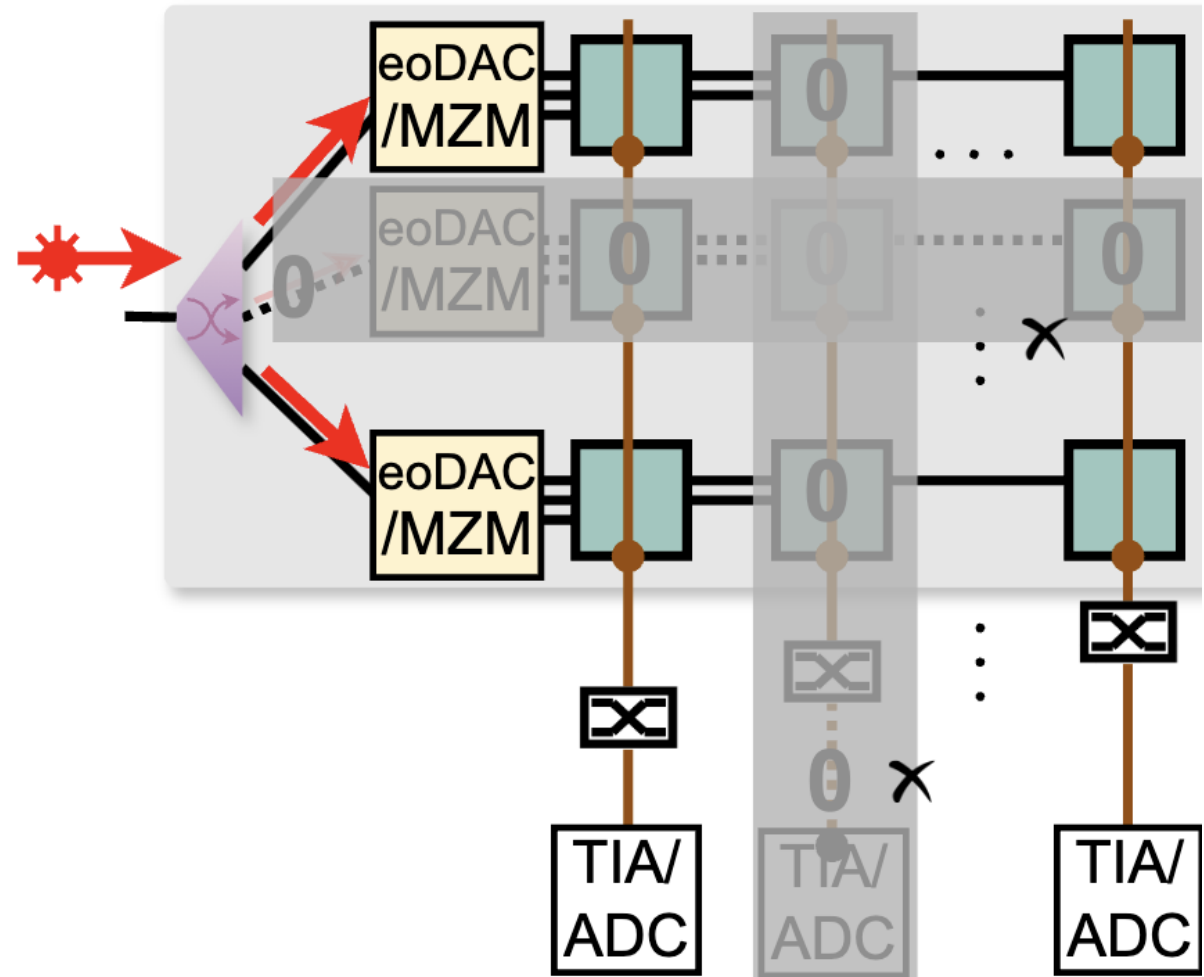
Proposed Co-Sparse Photonic Tensor Accelerator SCATTER



Proposed Reconfigurable Sparse Photonic Accel.

- ~~Lacks universality~~
- **Universal full-range PTC**
- Phase/Thermal sensitive
- Need large device spacing to reduce crosstalk
- Analog device/circuit noise
- Large on-chip area cost
- Power-consuming E-O conversion

**Proposed Co-Sparse Photonic
Tensor Accelerator SCATTER**

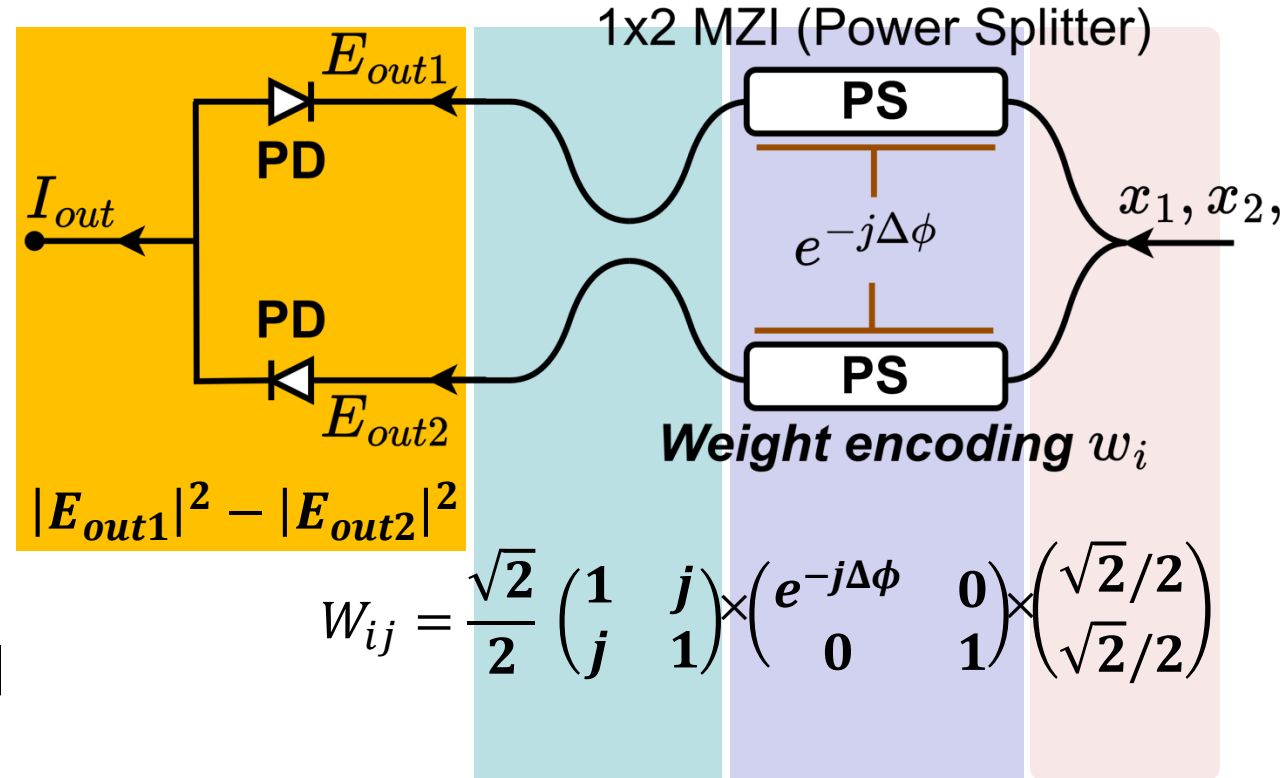


Universal Full Range Photonic Tensor Core

◆ Full-range dot-product engine

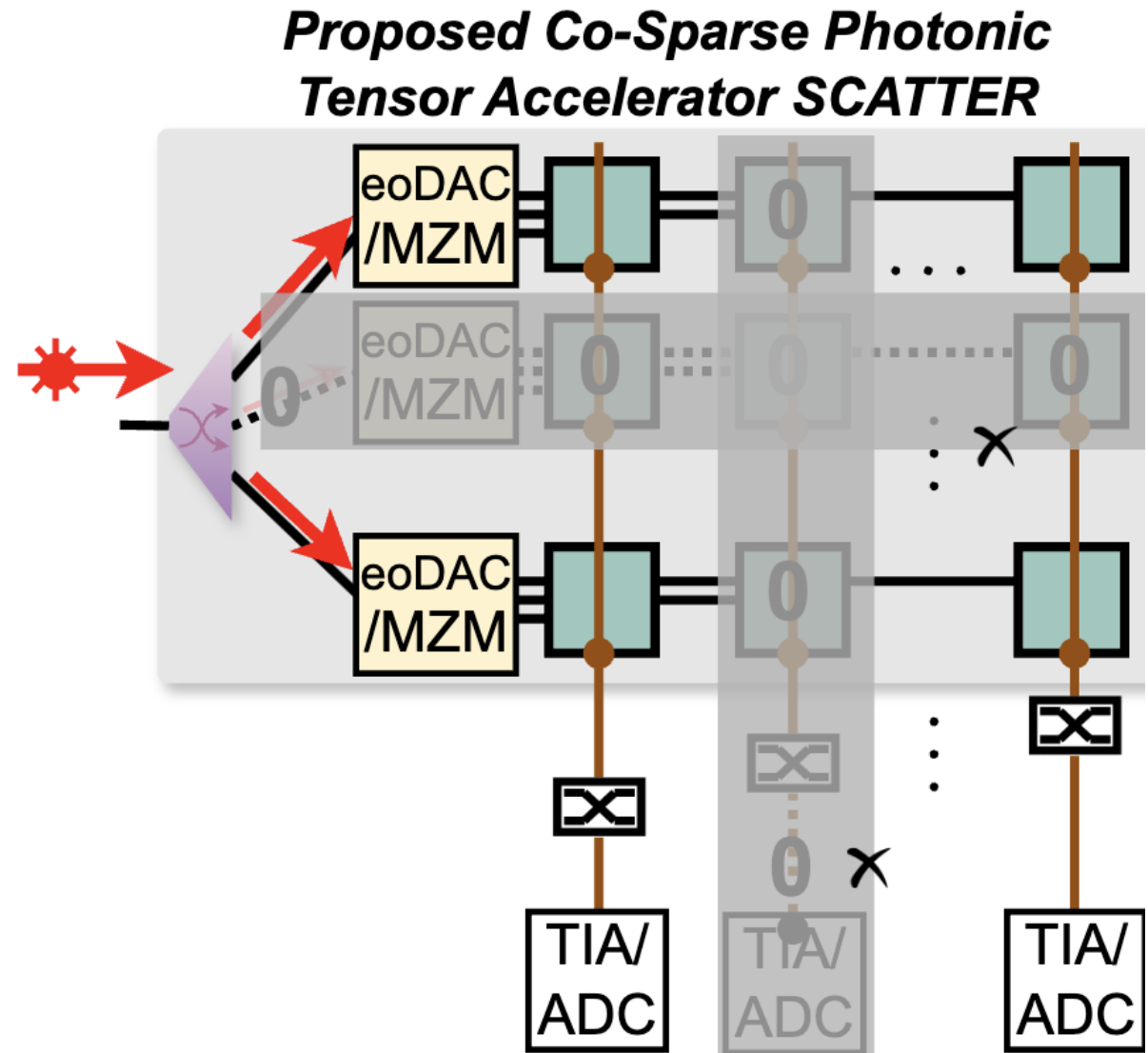
- > **1 × 2 MZI power splitter**
- > **Balanced** photodetectors enables **differential** output
- > Combining above two designs enables **full-range** weight representation

$$\begin{aligned} \text{» } I_{out} &= \underbrace{(\cos((\Delta\phi + \phi_b)))}_{W_{ij}} x; \phi_b = \frac{\pi}{2}, \\ &\Delta\phi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \\ \text{» } (\cos((\Delta\phi + \phi_b))) &\in [-1, 1] \end{aligned}$$



Proposed Reconfigurable Sparse Photonic Accel.

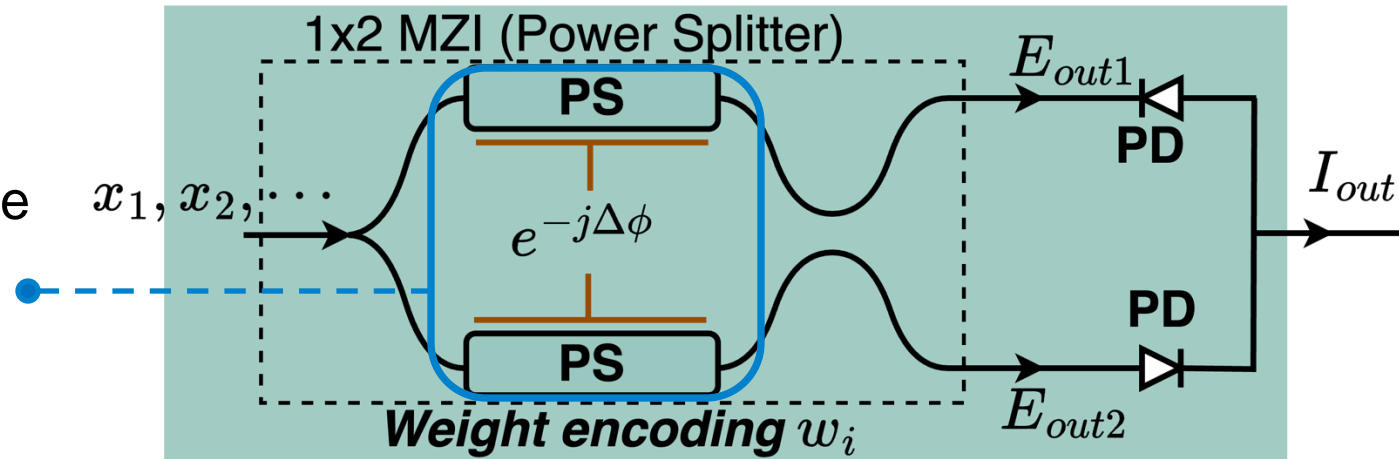
- ~~Lacks universality~~
- **Universal full-range PTC**
- ~~Phase/Thermal sensitive~~
- **Incoherent tensor core and symmetrical placement**
- Need large device spacing to reduce crosstalk
- Analog device/circuit noise
- Large on-chip area cost
- Power-consuming E-O conversion



Phase/Thermal-Insensitive Photonic Tensor Core

◆ Phase-Insensitive Design

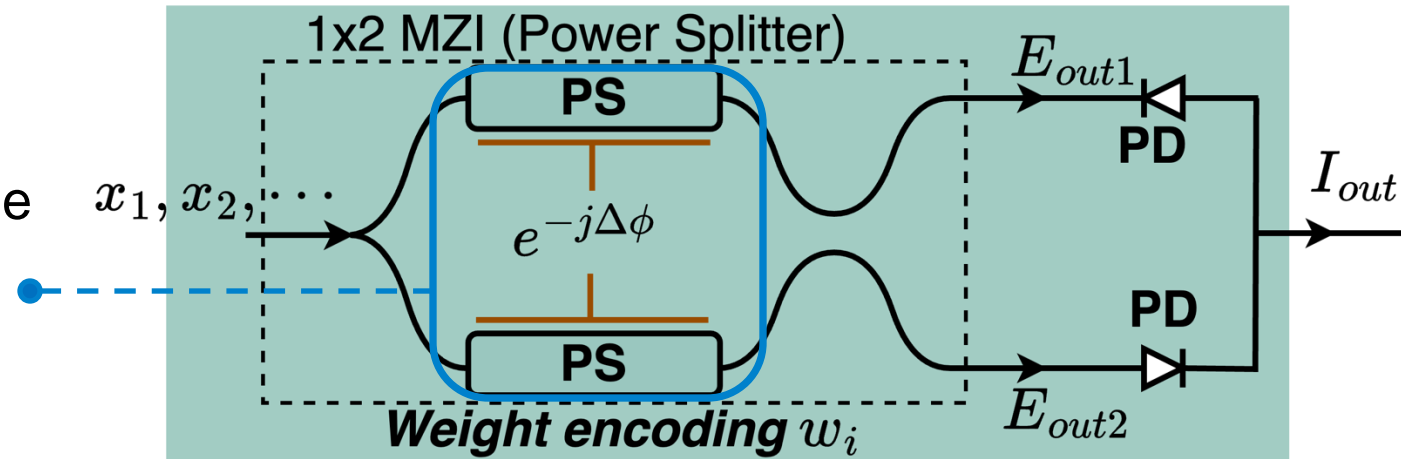
- › **Incoherent** tensor core
- › Use **power splitting ratio** to encode weight
- › **No phase** information used



Phase/Thermal-Insensitive Photonic Tensor Core

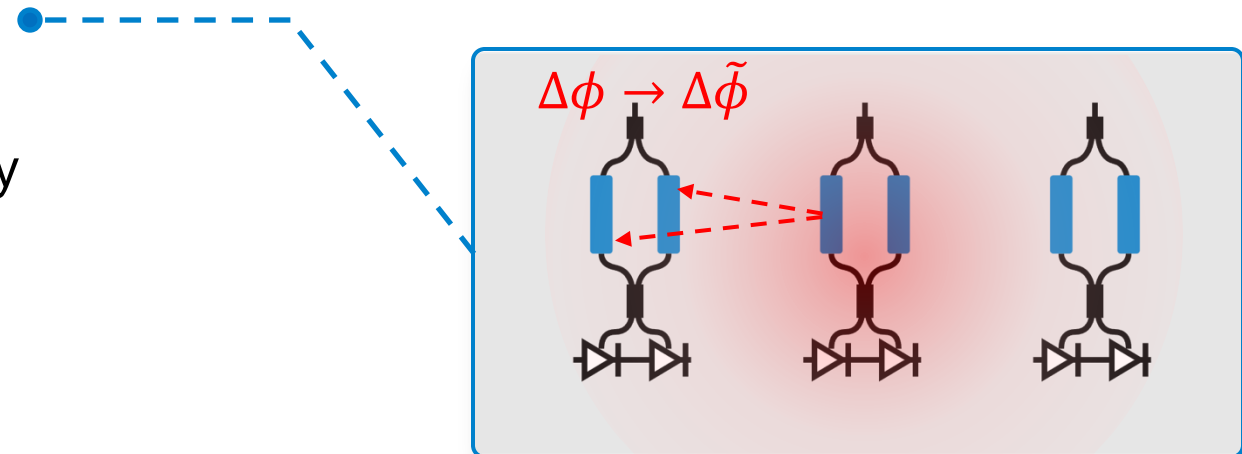
◆ Phase-Insensitive Design

- › **Incoherent** tensor core
- › Use **power splitting ratio** to encode weight
- › **No phase** information used



◆ Thermal-Insensitive Design

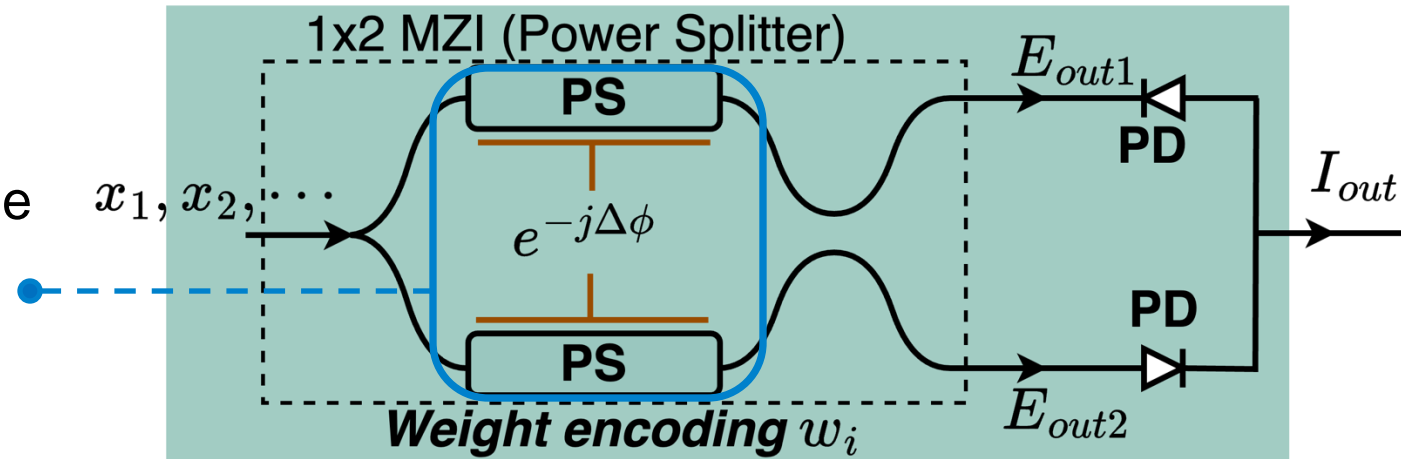
- › **Broadband** device
- › **Symmetrical** device placement
- › Phase error on two arms partially **cancel out**



Phase/Thermal-Insensitive Photonic Tensor Core

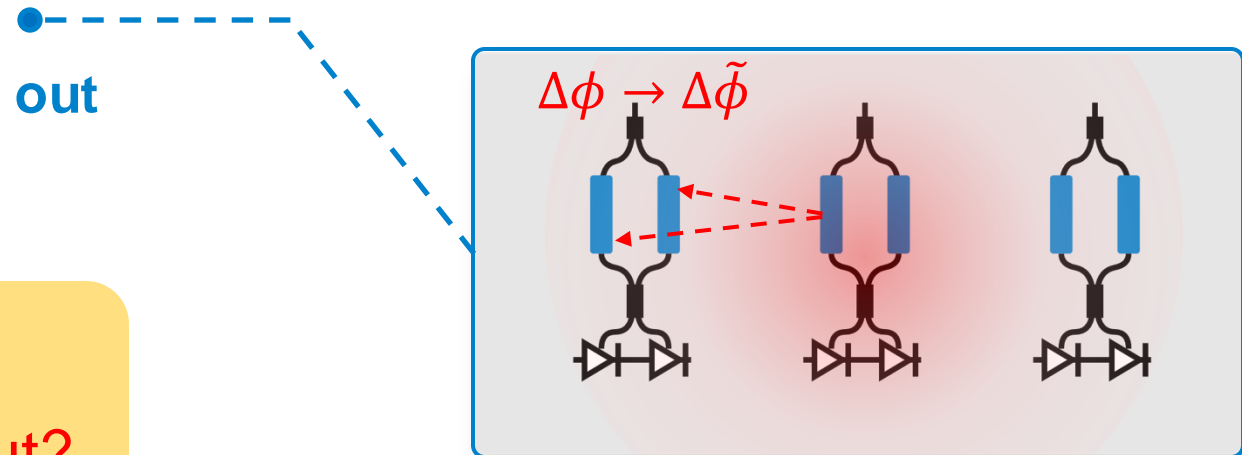
◆ Phase-Insensitive Design

- › **Incoherent** tensor core
- › Use **power splitting ratio** to encode weight
- › **No phase** information used



◆ Thermal-Insensitive Design

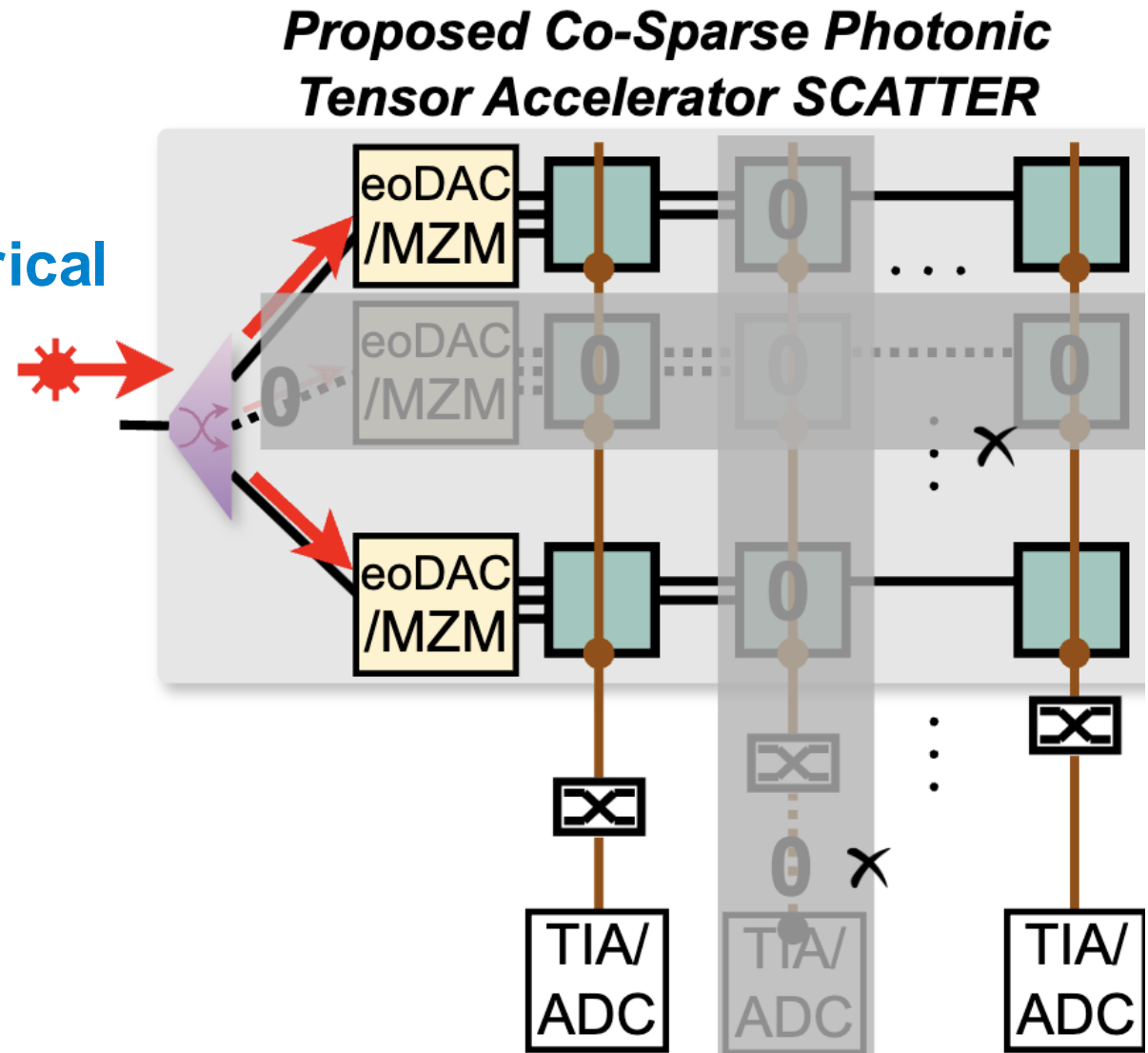
- › **Broadband** device
- › Phase error on two arms **cancel out**
- › **Symmetrical** device placement



How to further reduce the crosstalk with a compact layout?

Proposed Reconfigurable Sparse Photonic Accel.

- ~~Lacks universality~~
- **Universal full-range PTC**
- ~~Phase/Thermal sensitive~~
- **Incoherent tensor core and symmetrical placement**
- ~~Need large device spacing to reduce crosstalk~~
- **Row pruning and output gating**
- Analog device/circuit noise
- Large on-chip area cost
- Power-consuming E-O conversion



Row Pruning + Output Gating to Reduce Crosstalk

◆ Row-wise interleave structural pruning + output gating

- › Reduce the error induced by crosstalk
- › Save power of unused TIA/ADC

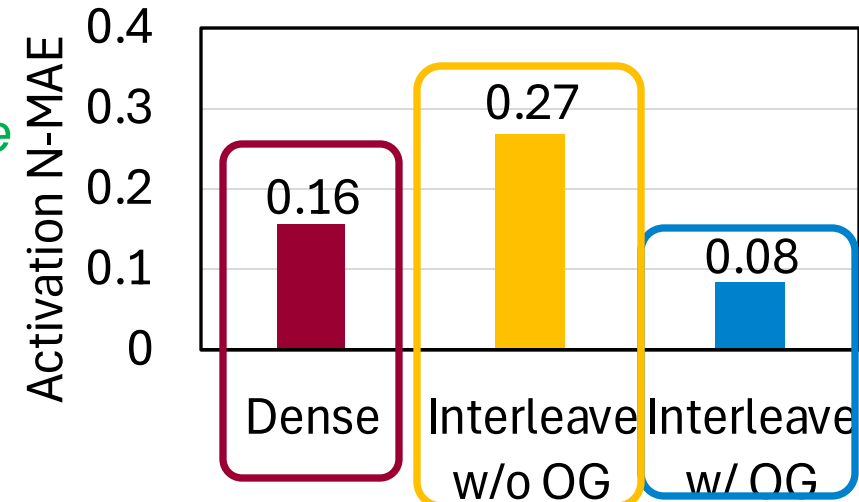
Ideal Weights

Dense

Interleaved prune

0.9	0.8
0.1	0.2
1.2	1.8

0.9	0.8
0	0
1.2	1.8

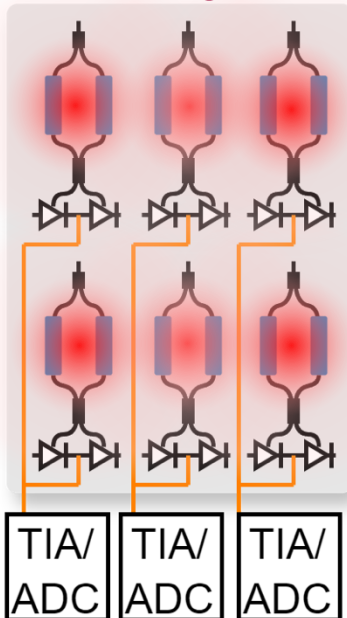


Real Weights

0.9	0.8
+0.24	+0.23
0.1	0.2
+0.12	+0.13
1.2	1.8
-0.25	-0.23

Dense

Spacing 5 μ m

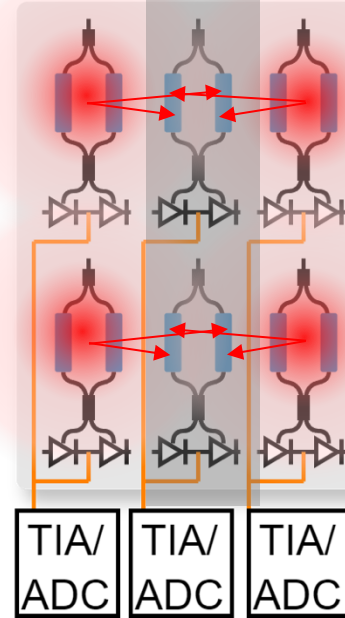


Interleave Pruning w/o OG

Real Weights

0.9	0.8
+0.06	+0.03
0	0
+0.46	+0.51
1.2	1.8
-0.07	-0.08

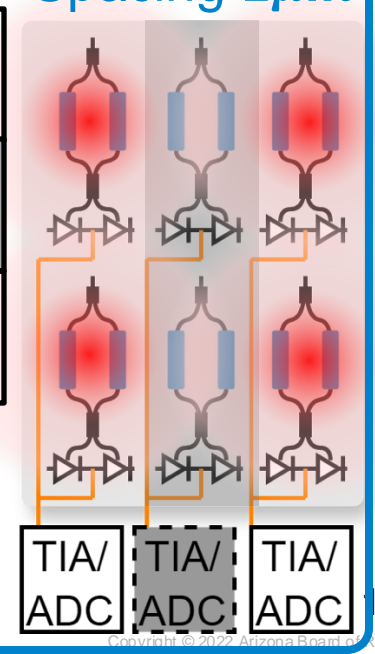
Spacing 5 μ m



Crosstalk induced error applies on non-operating MZI

Interleave Pruning w/ OG
Real Weights Spacing 1 μ m

0.9	0.8
+0.07	+0.04
0	0
+0.0	+0.0
1.2	1.8
-0.06	-0.09



OG eliminates error \rightarrow robustness \uparrow

Device tuning heat causes crosstalk to neighbor devices

Column Pruning + Input Gate ?

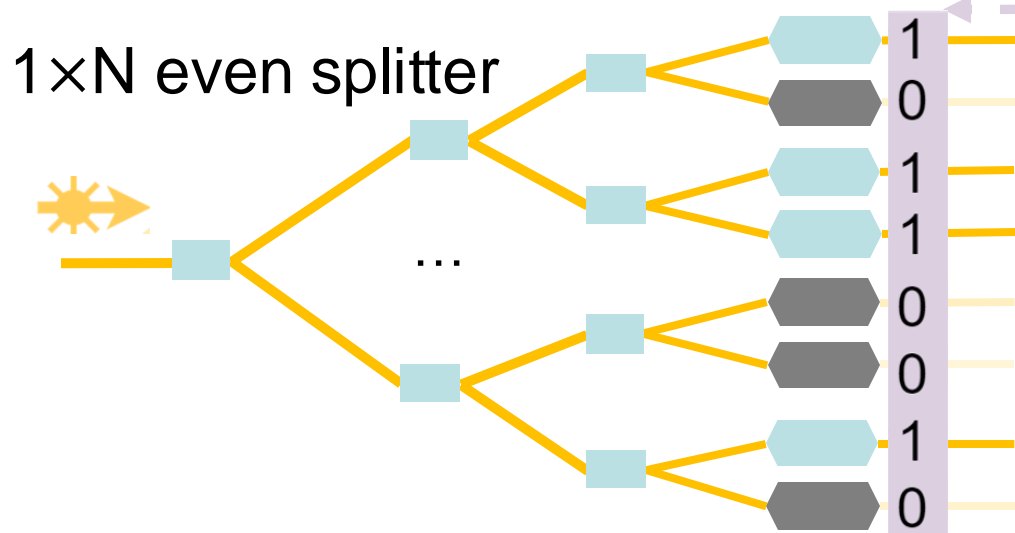
◆ Pruning w/ IG

Input leakage error Photodetector noise

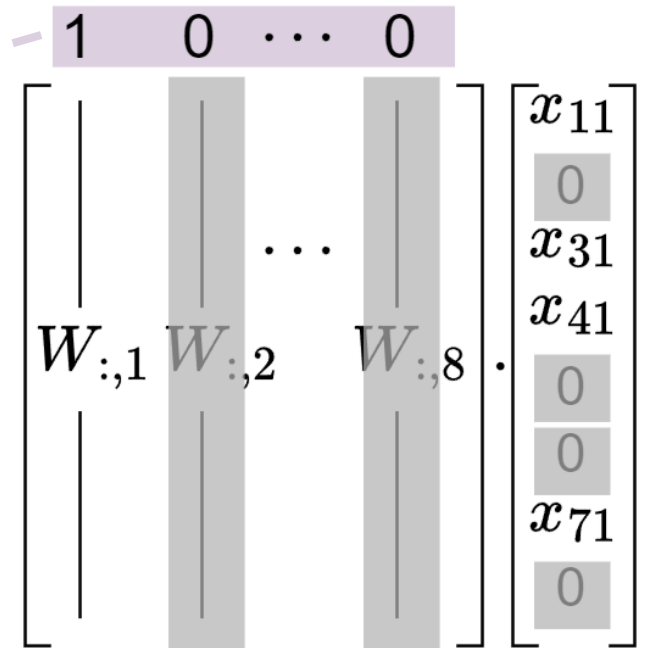
$$\text{> } y_{\text{prune}} = \sum_{j, m_j^c=1}^{k^2} (\widetilde{w}_j \cdot x_j) + \sum_{j, m_j^c=0}^{k^2} (\delta w_j \cdot \delta x_j) + 1 \cdot \sum_j^{k^2} (\delta n_{PD})$$

> Leakage error cannot be eliminated

How to eliminate leakage error?

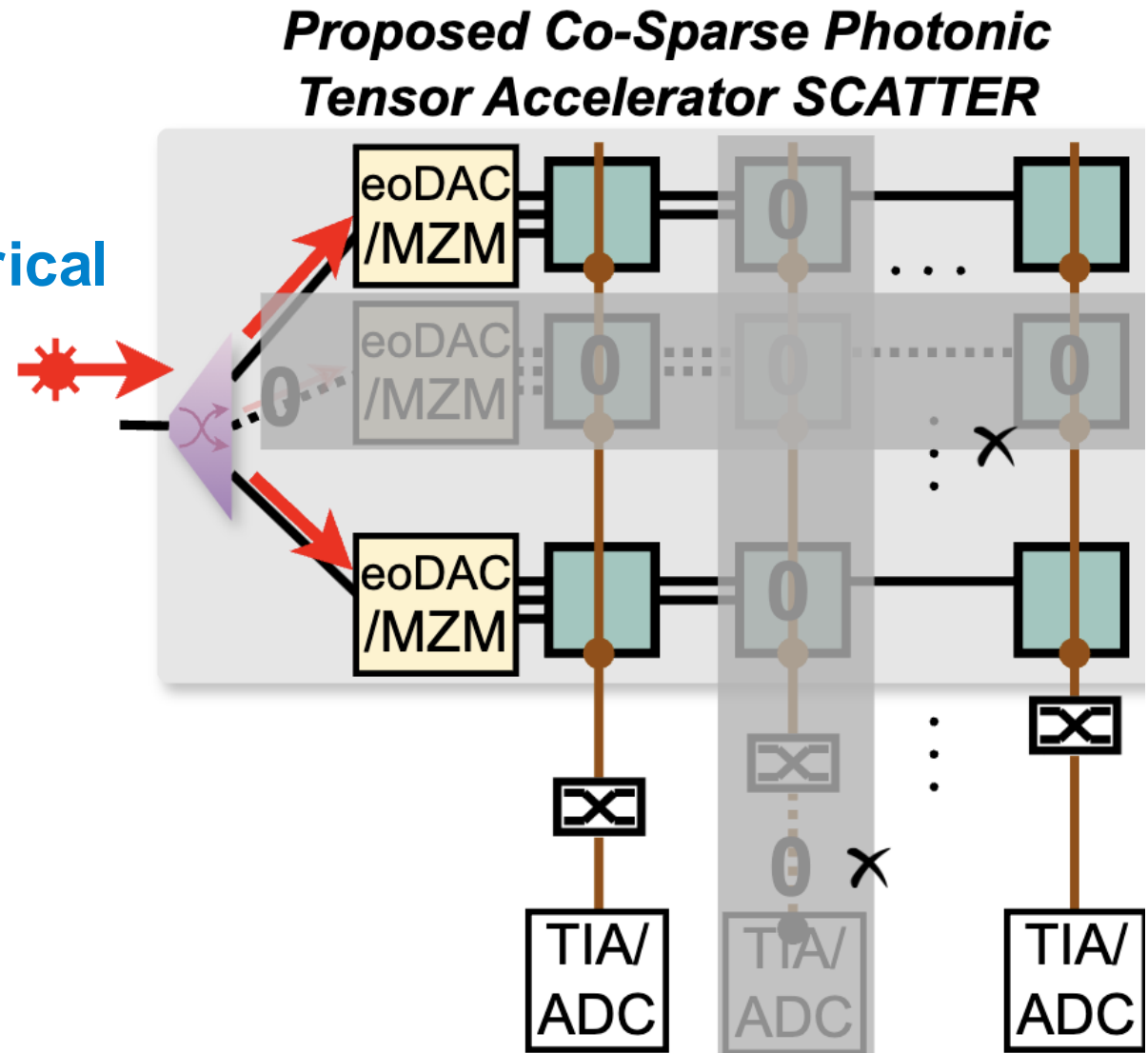


Column Sparsity Mask



Proposed Reconfigurable Sparse Photonic Accel.

- ~~Lacks universality~~
- **Universal full-range PTC**
- ~~Phase/Thermal sensitive~~
- **Incoherent tensor core and symmetrical placement**
- ~~Need large device spacing to reduce crosstalk~~
- **Row pruning and output gating**
- ~~Analog device/circuit noise~~
- **Light redistribution**
- Large on-chip area cost
- Power-consuming E-O conversion



Light Redis. to Reduce Leakage and PD Error

◆ Pruning w/ IG

Input leakage error Photodetector noise

$$\text{> } y_{\text{prune}} = \sum_{j, m_j^c=1}^{k^2} (\widetilde{w}_j \cdot x_j) + \sum_{j, m_j^c=0}^{k^2} (\delta w_j \cdot \delta x_j) + 1 \cdot \sum_j^{k^2} (\delta n_{PD})$$

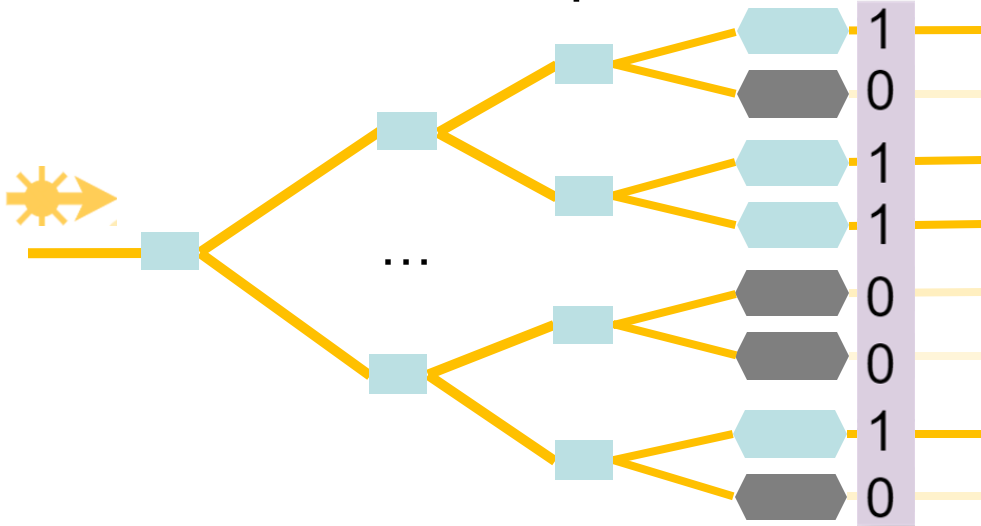
◆ Pruning w/ IG and light redistribution

$$\text{> } y_{\text{prune}} = \sum_{j, m_j^c=1}^{k^2} (\widetilde{w}_j \cdot x_j) + 0 + \frac{k'_2}{k_2} \sum_j^{k^2} (\delta n_{PD})$$

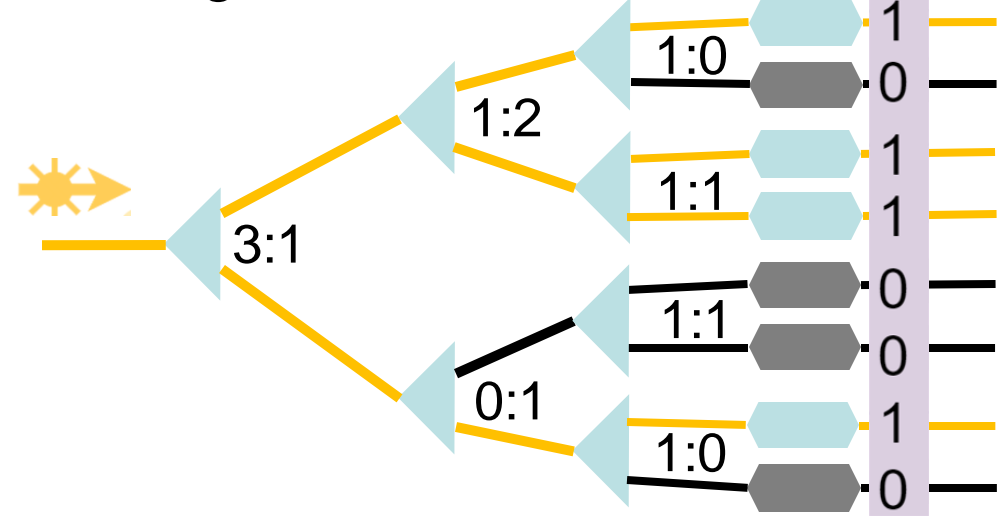
No leakage error

Save light power
Higher SNR

1×N even splitter

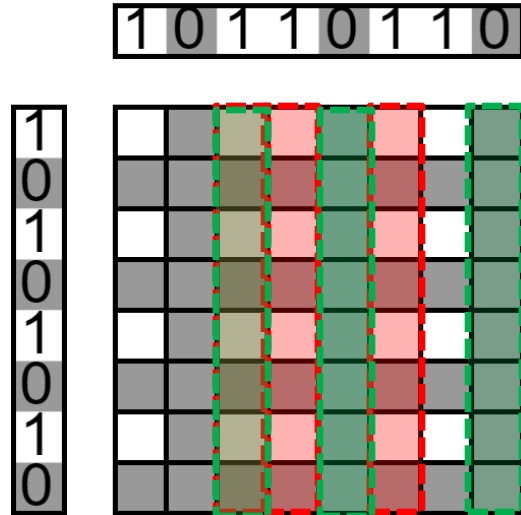


In-situ light rerouter with redistribution

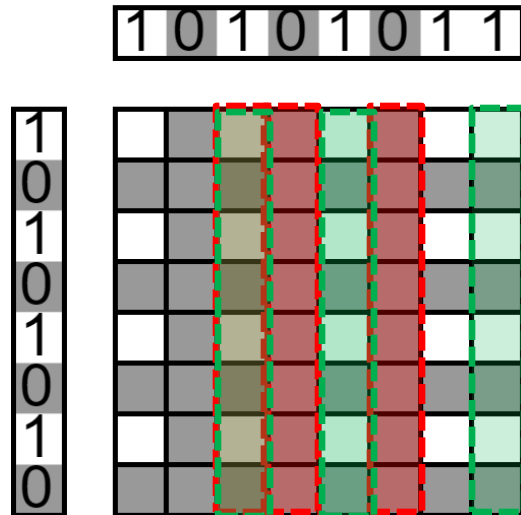


Crosstalk and Power-Aware Dynamic Sparse Training

Initialize the sparsity mask



Row-sparsity mask remains interleaved shape



Final sparsity mask

Prune stage
Which to *prune*? `1 0 0 0 0 0 1 0`

1. Small column norm
2. Minimize MZI crosstalk
3. Minimize rerouter power



Iteratively explore high-accuracy, robust, and efficient sparsity pattern



Growth stage
Which to *grow*? `1 0 1 0 1 0 1 1`

1. Large column gradient norm
2. Minimize MZI crosstalk
3. Minimize rerouter power

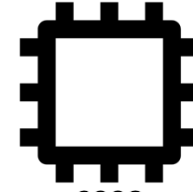
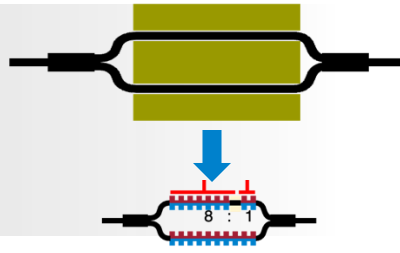


Cross-Stack Power/Area/Robustness Optimization

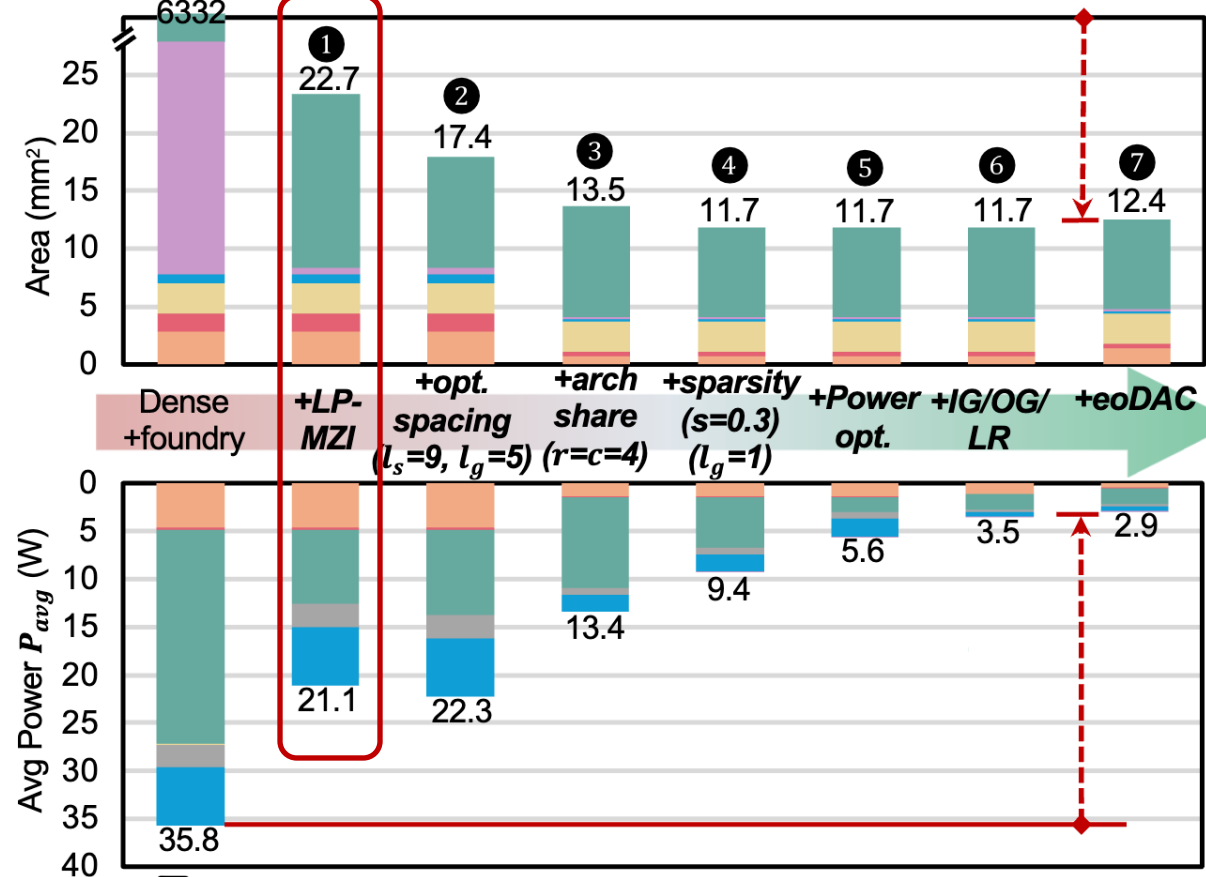
1

✓ Customized low-power device

~280× smaller + 1.7× power ↓



511× more compact



12.4× more efficient



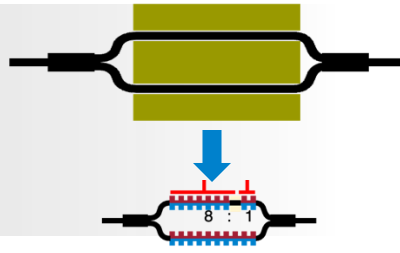
Z. Yin, N. Gangi, M. Zhang, J. Zhang, R. Huang, J. Gu, "SCATTER: Algorithm-Circuit Co-Sparse Photonic Accelerator with Thermal-Tolerant, Power-Efficient In-situ Light Redistribution," *ACM/IEEE ICCAD*, 2024.

Cross-Stack Power/Area/Robustness Optimization

1

✓ Customized low-power device

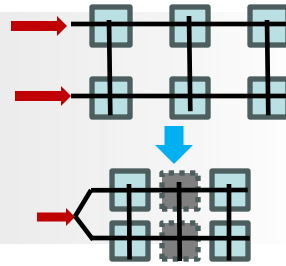
~280× smaller + 1.7× power ↓



2

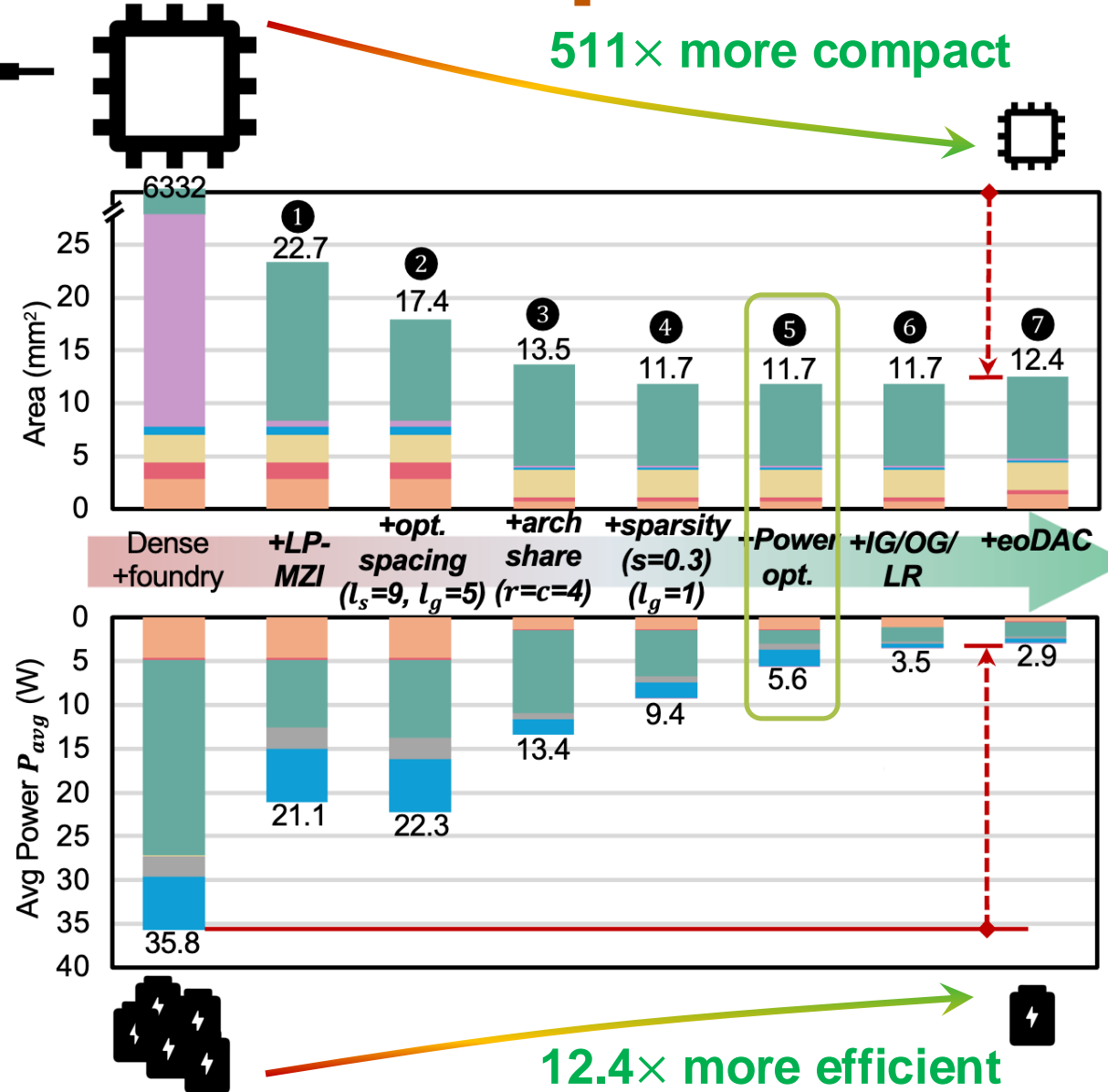
✓ Optimized layout + Sparse arch

1.5× smaller + 4× power ↓



5

Power/crosstalk-aware optimization



511× more compact

12.4× more efficient

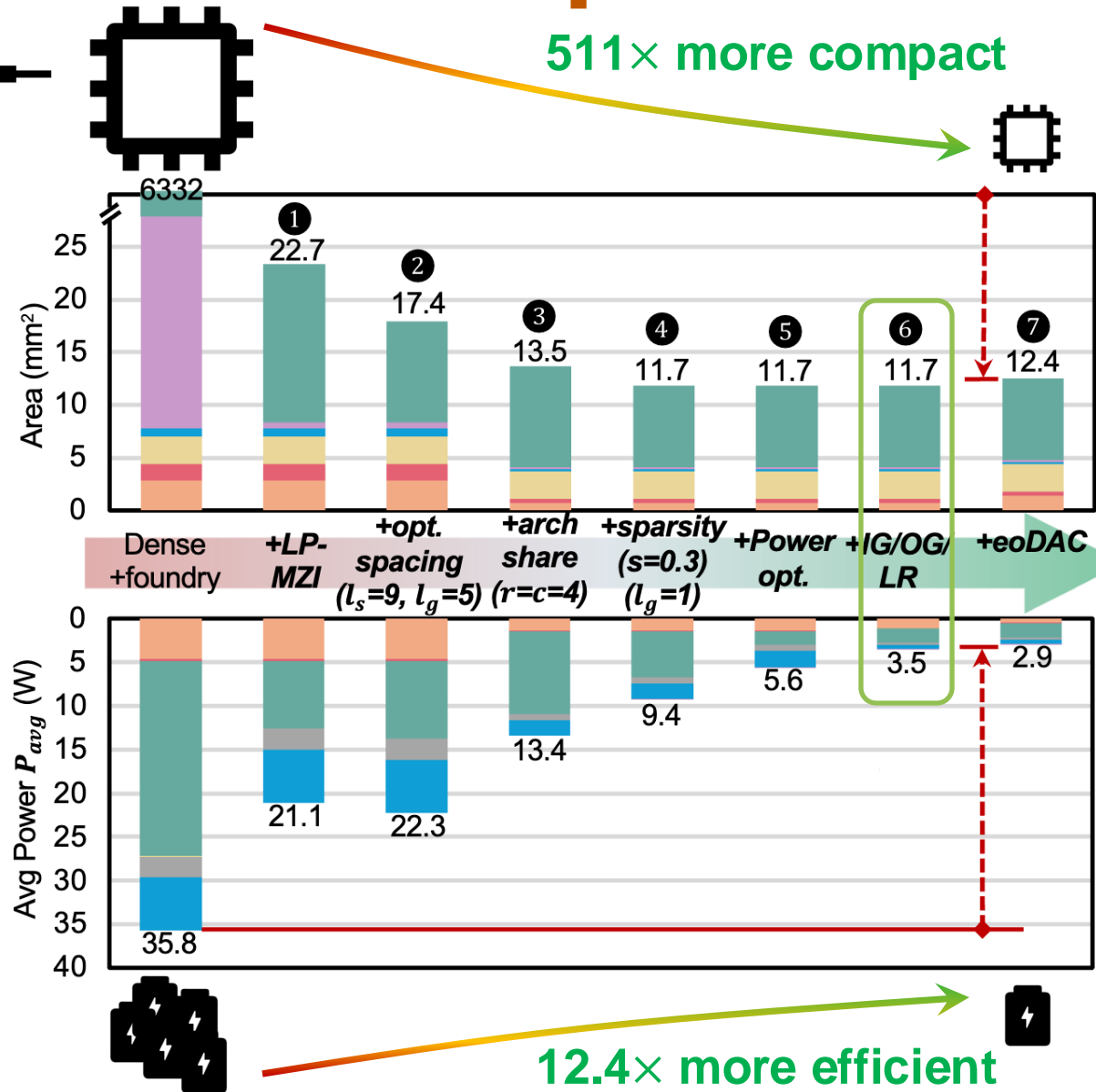
Cross-Stack Power/Area/Robustness Optimization

1 ✓ **Customized low-power device**
 ~280× smaller + 1.7× power ↓

2 ✓ **Optimized layout + Sparse arch**
 1.5× smaller + 4× power ↓

5 Power/crosstalk-aware optimization

6 ✓ **Light redistribution + gating**
 No crosstalk + 1.6× power ↓
 Reconfigurable sparse matrix multiply



Cross-Stack Power/Area/Robustness Optimization

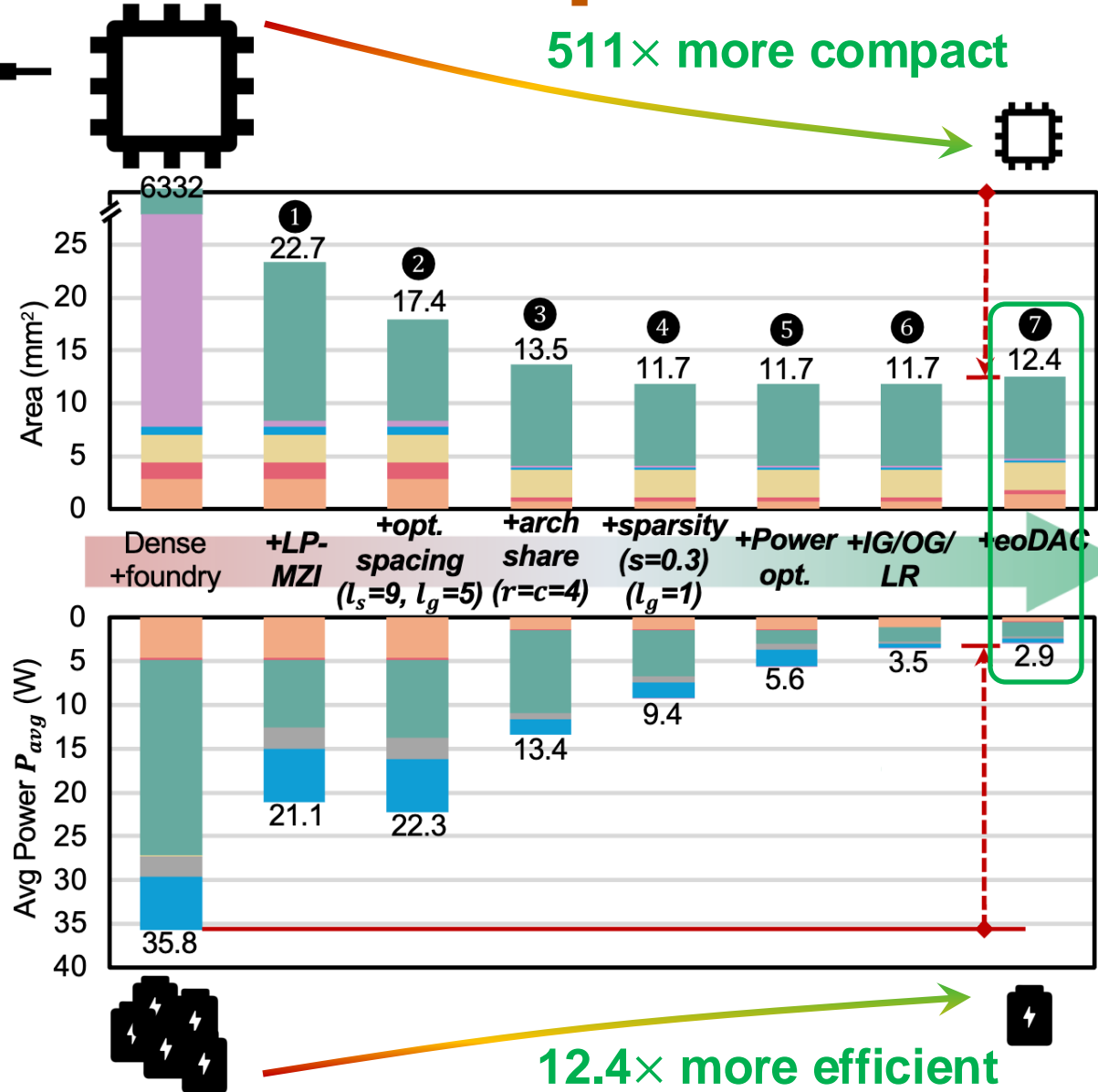
1 ✓ **Customized low-power device**
 ~280× smaller + 1.7× power ↓

2 ✓ **Optimized layout + Sparse arch**
 1.5× smaller + 4× power ↓

5 Power/crosstalk-aware optimization

6 ✓ **Light redistribution + gating**
 No crosstalk + 1.6× power ↓
 Reconfigurable sparse matrix multiply

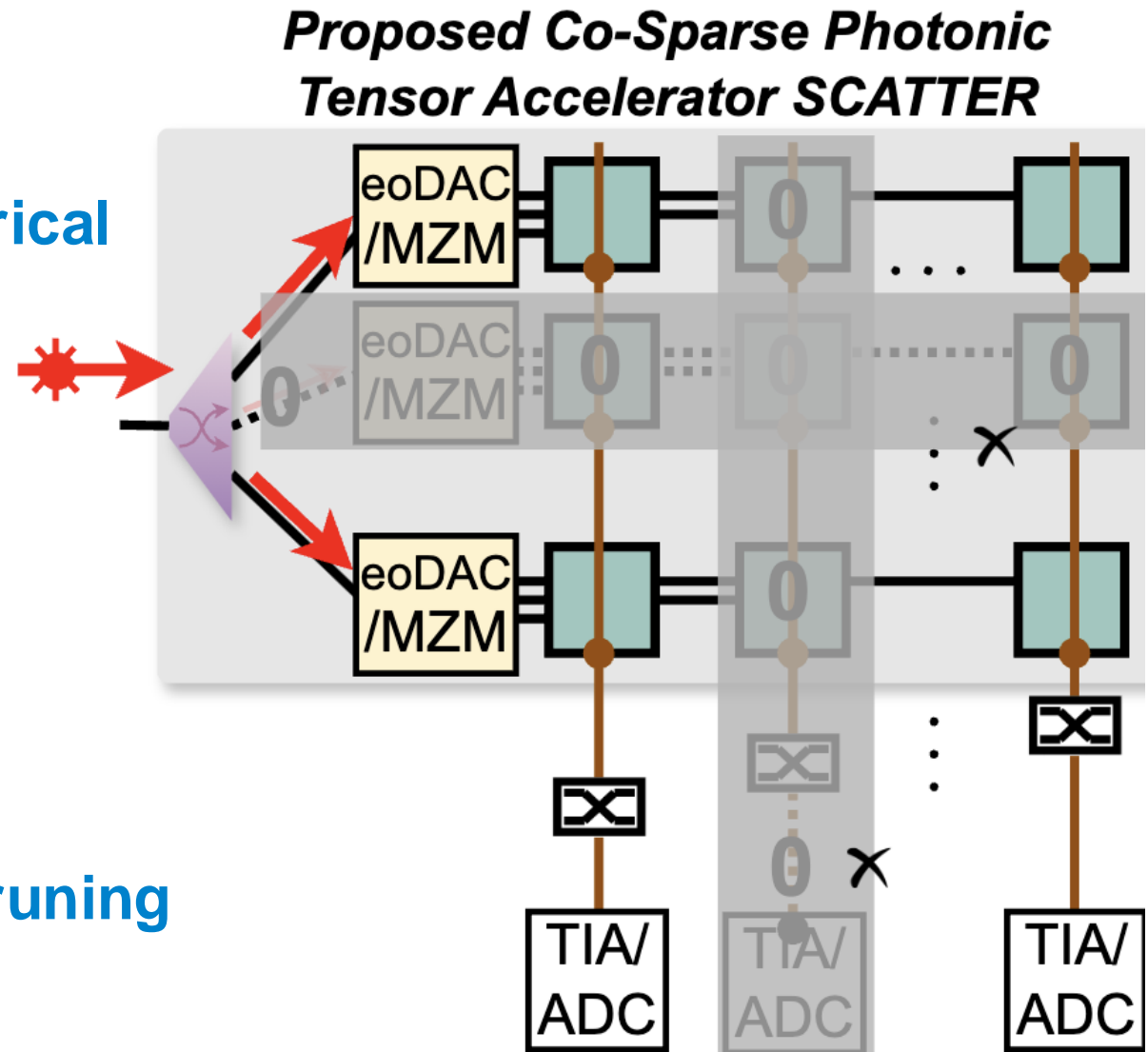
7 ✓ **Hybrid electronic-optical digital-to-analog converter**
 1.2× power ↓ + high precision



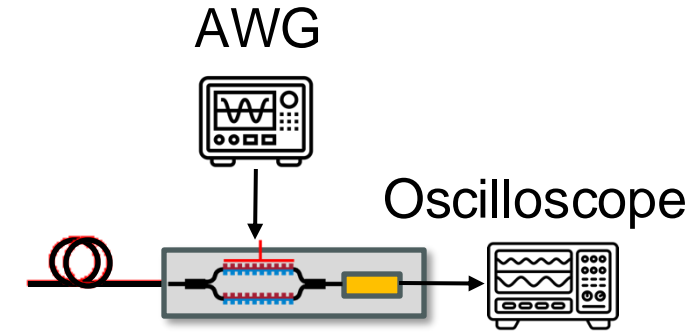
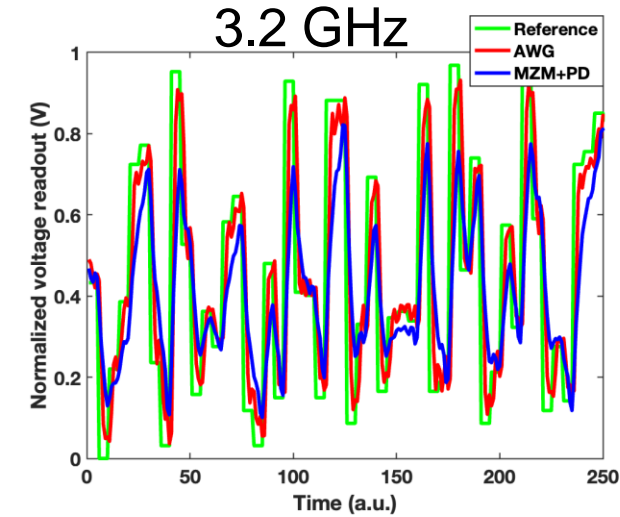
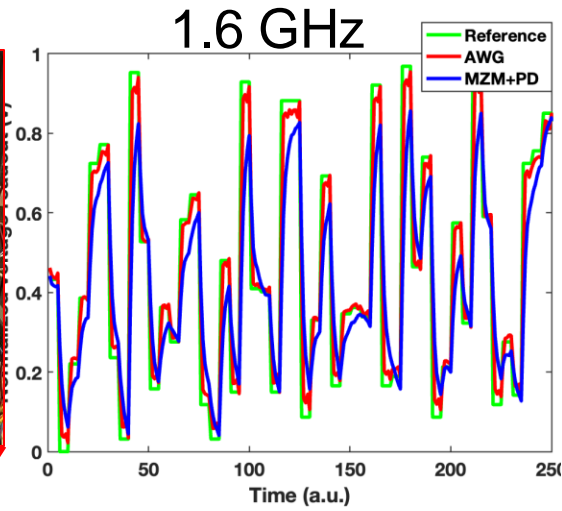
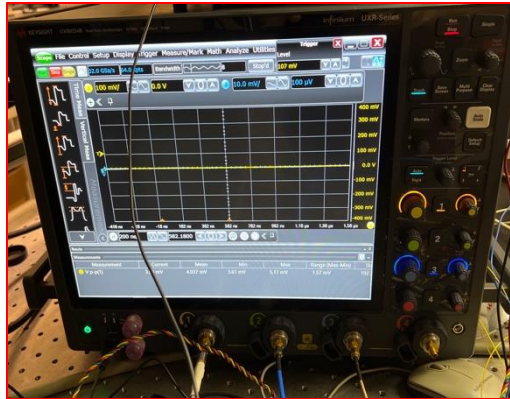
Z. Yin, N. Gangi, M. Zhang, J. Zhang, R. Huang, J. Gu, "SCATTER: Algorithm-Circuit Co-Sparse Photonic Accelerator with Thermal-Tolerant, Power-Efficient In-situ Light Redistribution," **ACM/IEEE ICCAD**, 2024.

Proposed Reconfigurable Sparse Photonic Accel.

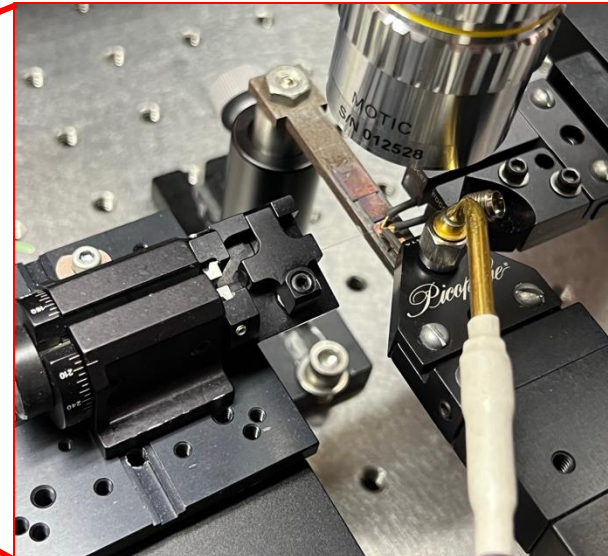
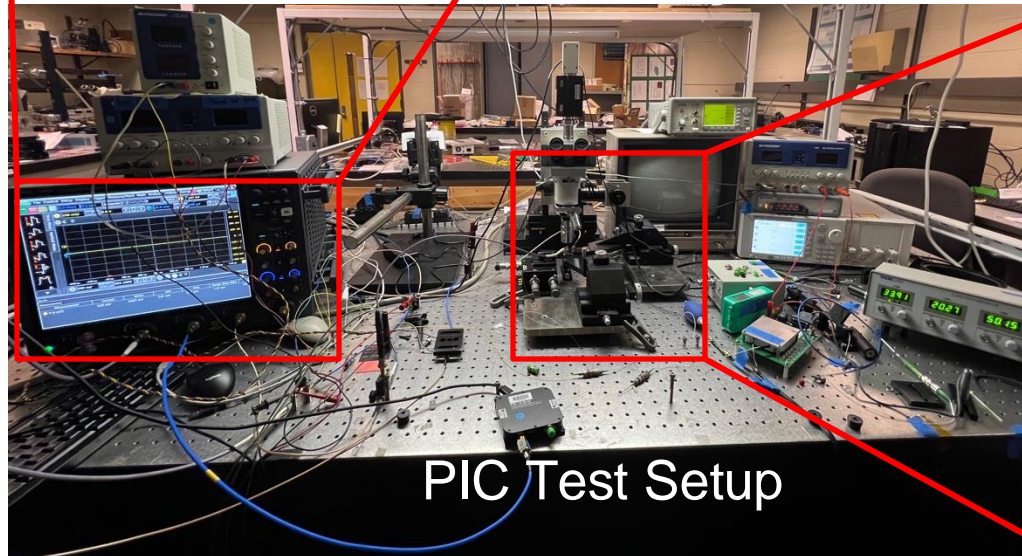
- ~~▪ Lacks universality~~
- **Universal full-range PTC**
- ~~▪ Phase/Thermal sensitive~~
- **Incoherent tensor core and symmetrical placement**
- ~~▪ Need large device spacing to reduce crosstalk~~
- **Row pruning and output gating**
- ~~▪ Analog device/circuit noise~~
- **Light redistribution and input gating**
- ~~▪ Large on-chip area cost~~
- **Customized device and structural pruning**
- ~~▪ Power-consuming E-O conversion~~
- **Device gating and hybrid eoDAC**



Single-Link Chip Testing and Error Calibration



Testing vs. Simulation:
0.8~3% relative square error
6~7-bit Resolution



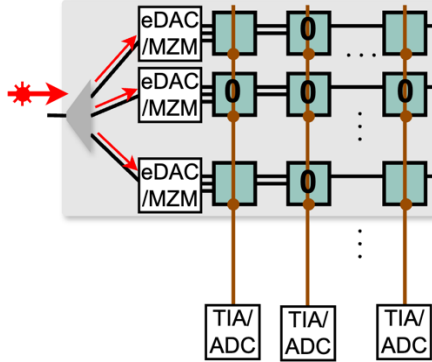
[Courtesy: Prof. Rena Huang's group for testing]

Application Evaluation and Efficiency Comparison

ResNet18 – CIFAR100

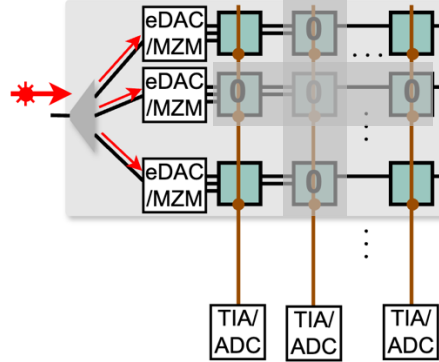
w/ Thermal Variation MZI Spacing $1\mu\text{m}$

Dense



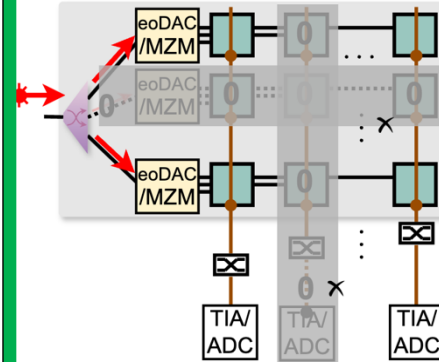
44.12%

Prune



0.51%

SCATTER

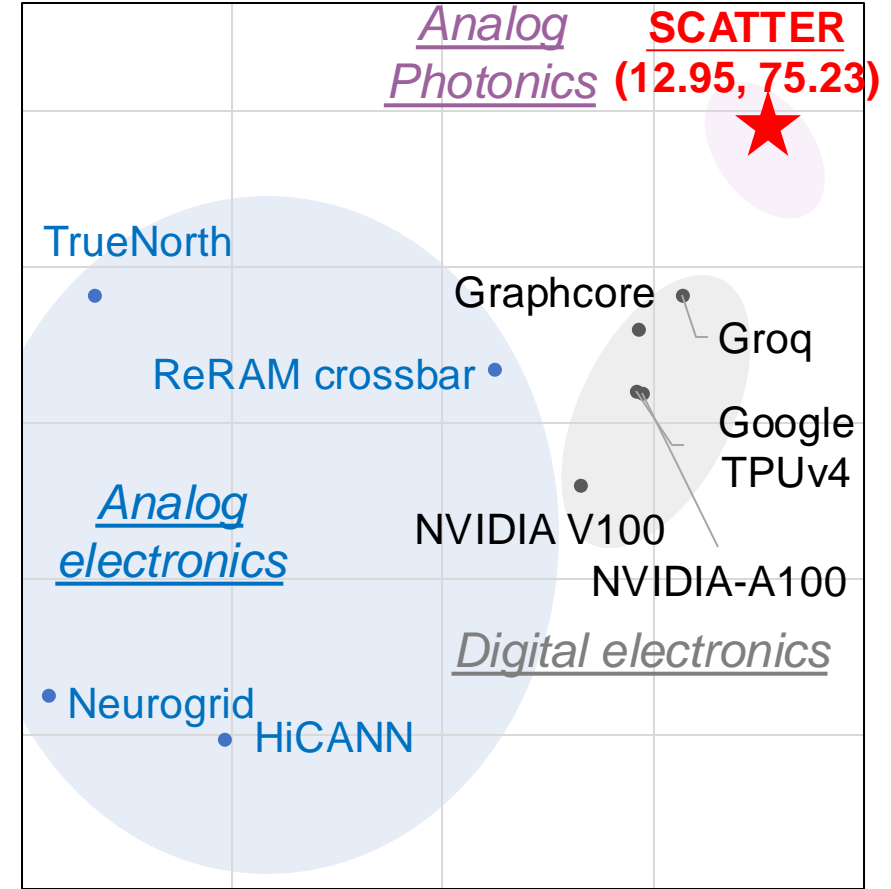


57.40%

Increasing the robustness against crosstalk and noise while maintaining a compact size

Energy Efficiency (TOPS/W)

1.E+02
1.E+01
1.E+00
1.E-01
1.E-02
1.E-03



Compute Density (TOPS/mm²)



Thank you!

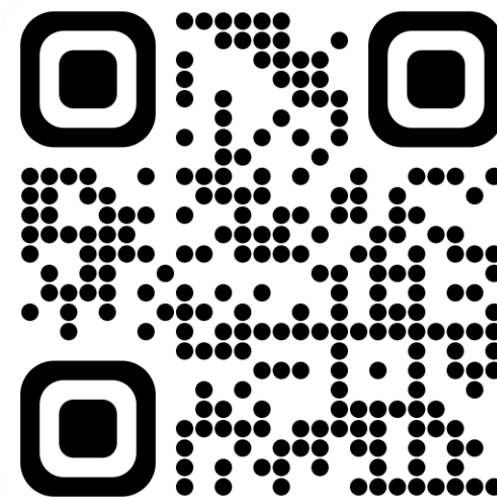
Q & A?



Open-Source
TorchONN Toolchain

SCATTER: Algorithm-Circuit Co-Sparse Photonic Accelerator with
Thermal-Tolerant, Power-Efficient In-situ Light Redistribution

Ziang Yin¹, Nicholas Gangi², Meng Zhang², Jeff Zhang¹, Rena Huang², Jiaqi Gu^{1†}
¹Arizona State University, ²Rensselaer Polytechnic Institute



arXiv Preprint

*Automating optical AI hardware
design toward productivity*



Rensselaer