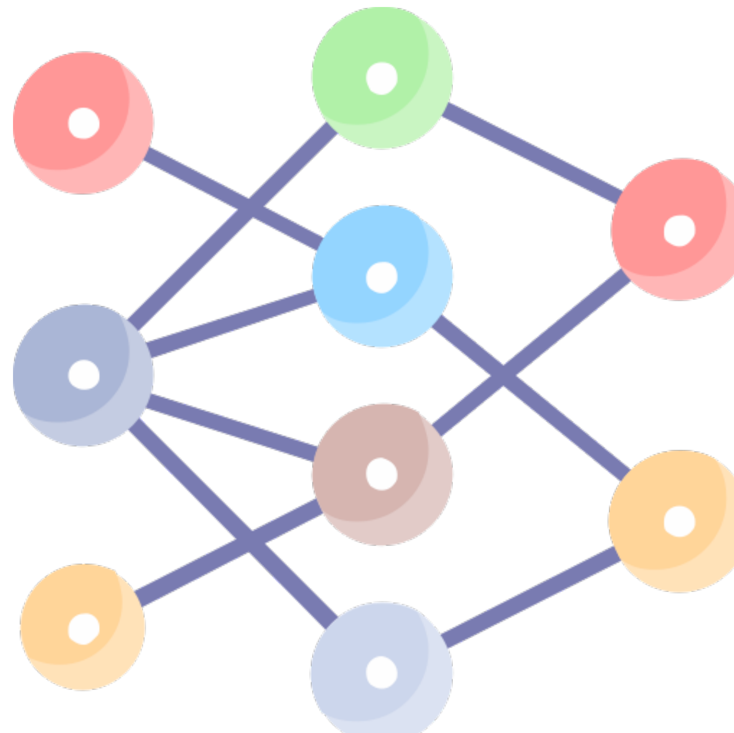
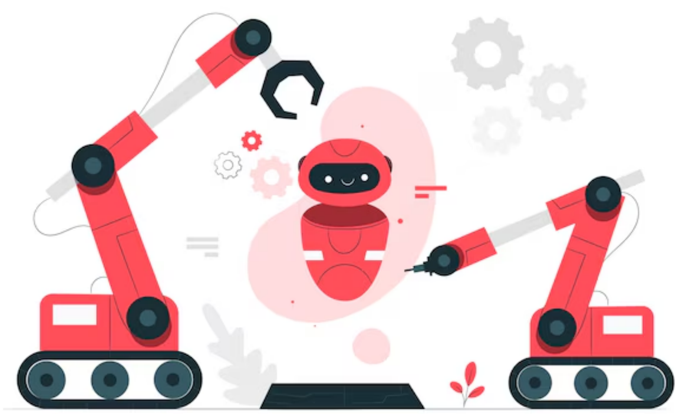
An Efficient General-Purpose Optical Accelerator for Neural Networks

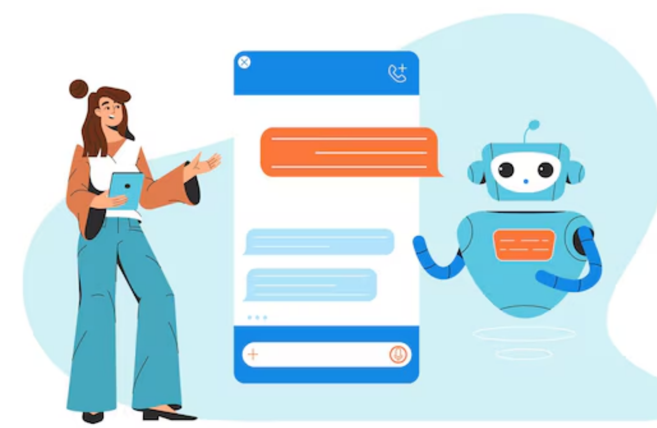
ASP-DAC 2025

Sijie Fei (Technical University of Munich)
Amro Eldebiky (Technical University of Munich)
Grace Li Zhang (Technical University of Darmstadt)
Bing Li (University of Siegen)
Ulf Schlichtmann (Technical University of Munich)

Why **O**ptical **N**eural **N**etworks (**ONN**)







- Strict time constraints
- Millions/Trillions Parameters

...

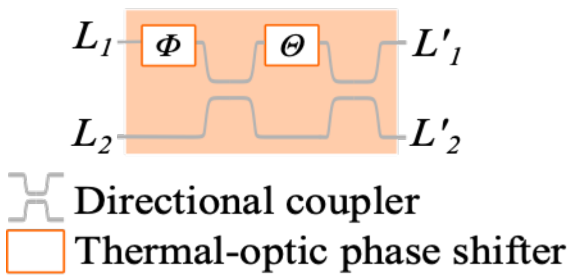
higher speed

ONN

Basics of ONN

Architecture dependent on matrix dimension

- **Mach-Zehnder Interferometers** (MZI)



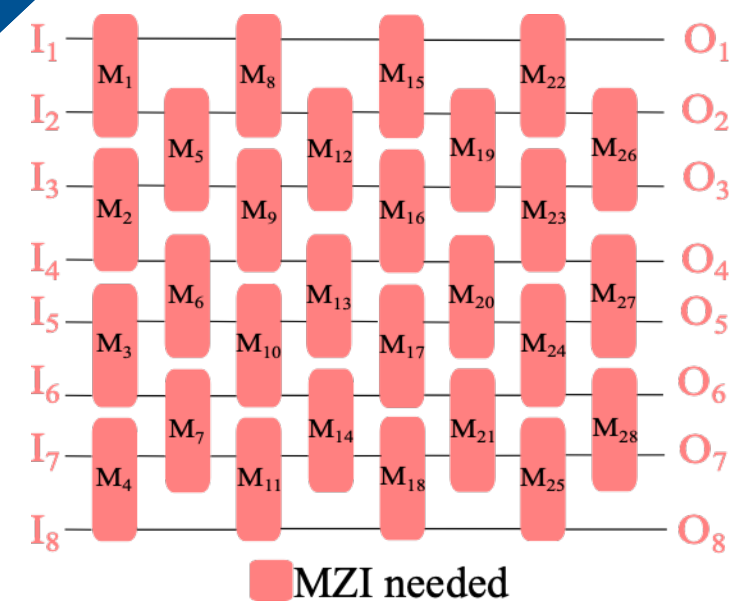
$$\begin{bmatrix} L'^{c}_1 \\ L'^{c}_2 \end{bmatrix} = ie^{\frac{i\theta}{2}} \begin{bmatrix} e^{i\phi}\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \\ e^{i\phi}\cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \end{bmatrix} \begin{bmatrix} L^{c}_1 \\ L^{c}_2 \end{bmatrix} = \mathbf{T} \begin{bmatrix} L^{c}_1 \\ L^{c}_2 \end{bmatrix}$$

Unitary

SVD

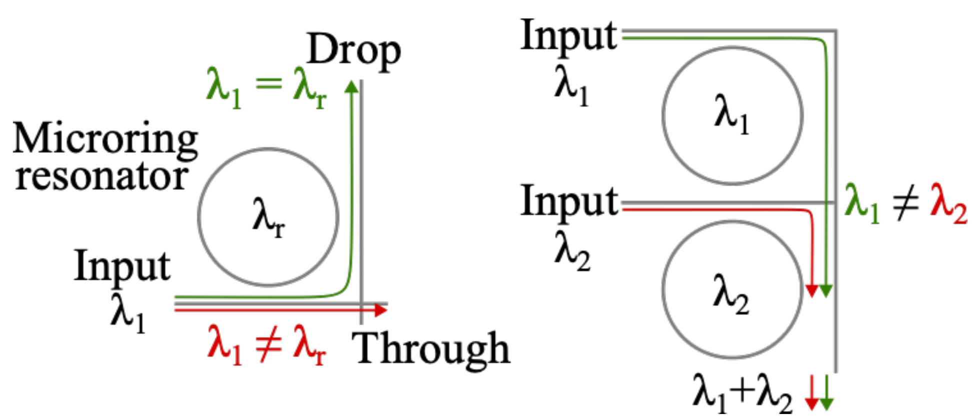
[6]

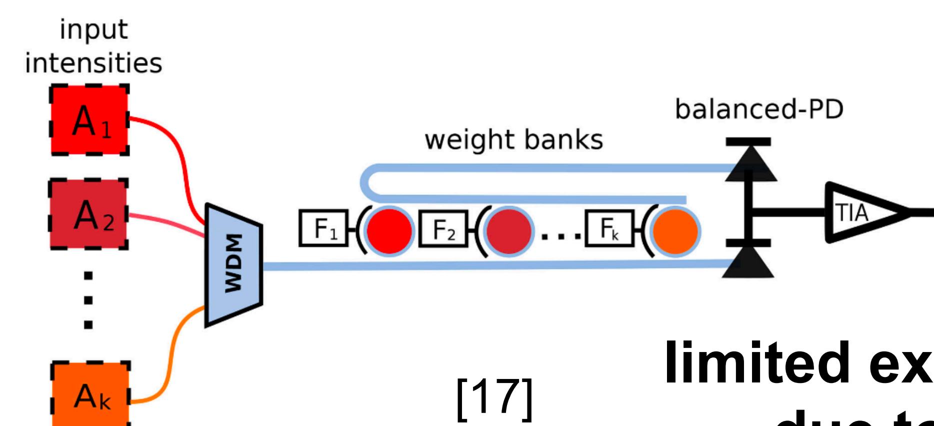
Weight Matrices with cascading MZIs



MxN Matrix ~ (M^2+N^2)/2 MZIs

- **Microring Resonator** (MRR)





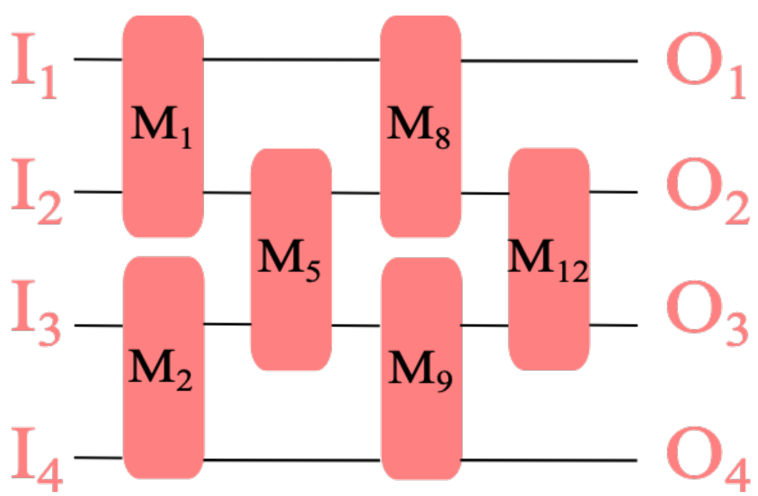
[17]

limited expressivity due to wdm

[6] YichenShen, NicholasCHarris, Skirlo, etal. Deep learning with coherent nanophotonic circuits. Nature photonics, 11(7):441–446, 2017.
[17] Viraj Bangari, Bicky A Marquez, Heidi Miller, et al. Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs). IEEE Journal of Selected Topics in Quantum Electronics, 26(1):1–13, 2019.

Limitation of **G**eneral-purpose **O**ptical **A**ccelerators(**GOA**)

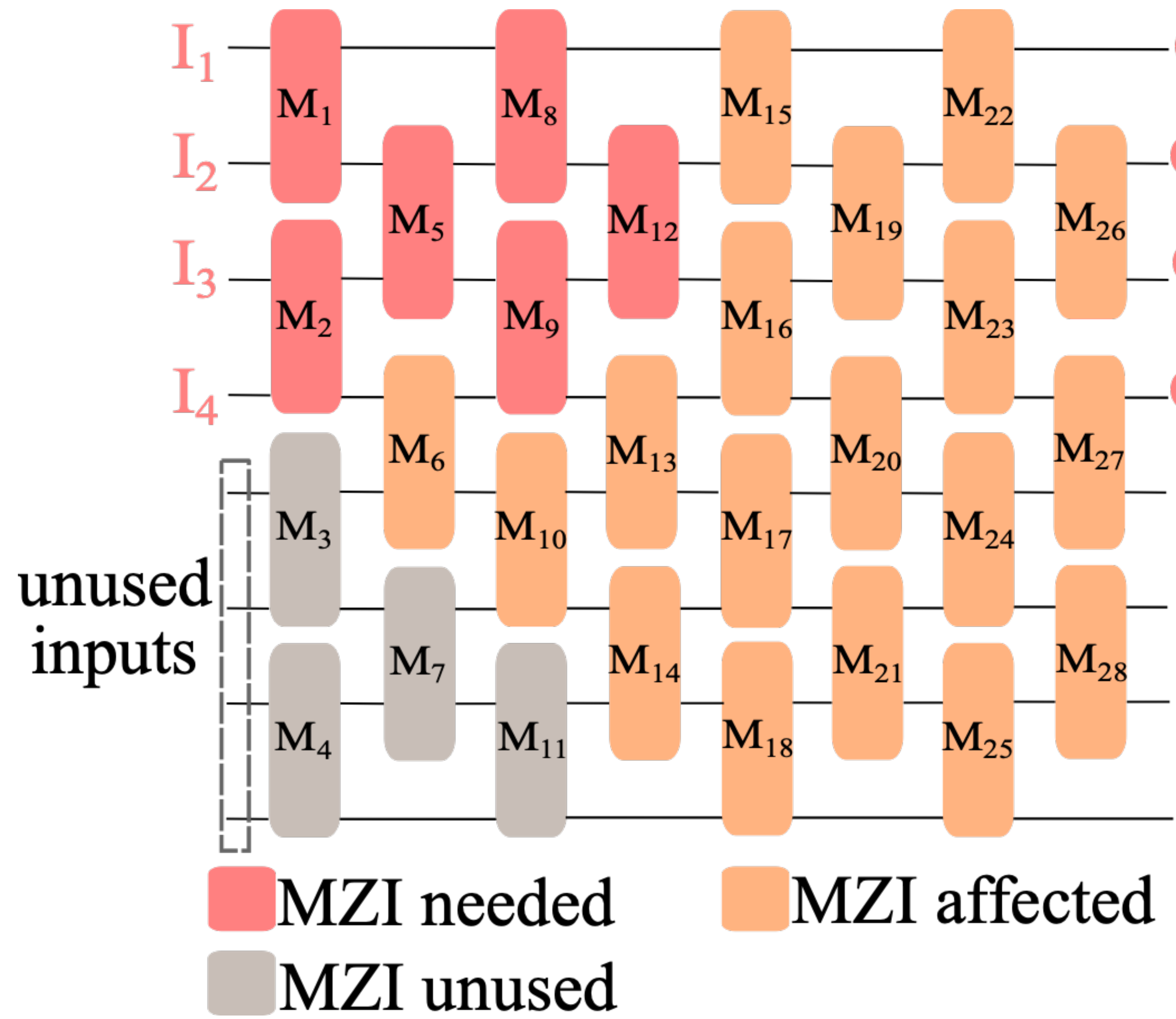
- **GOA:** the same optical accelerator that can be reused for different neural networks



a 4x4 matrix

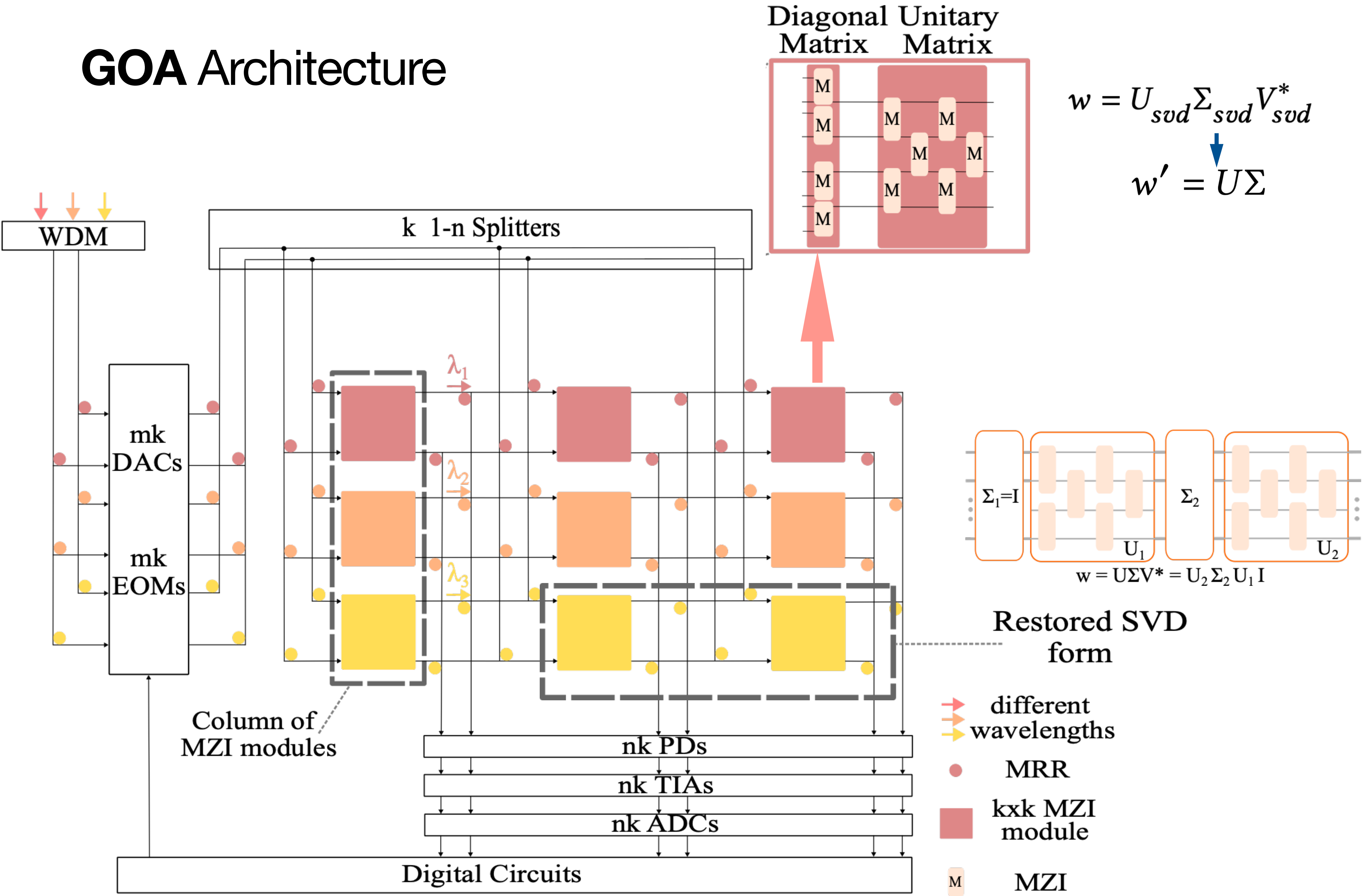
Mismatch of the matrix dimensions and GOA dimensions

Low utilization efficiency
High mapping effort
(Mapping: tuning PSs on GOA one time)



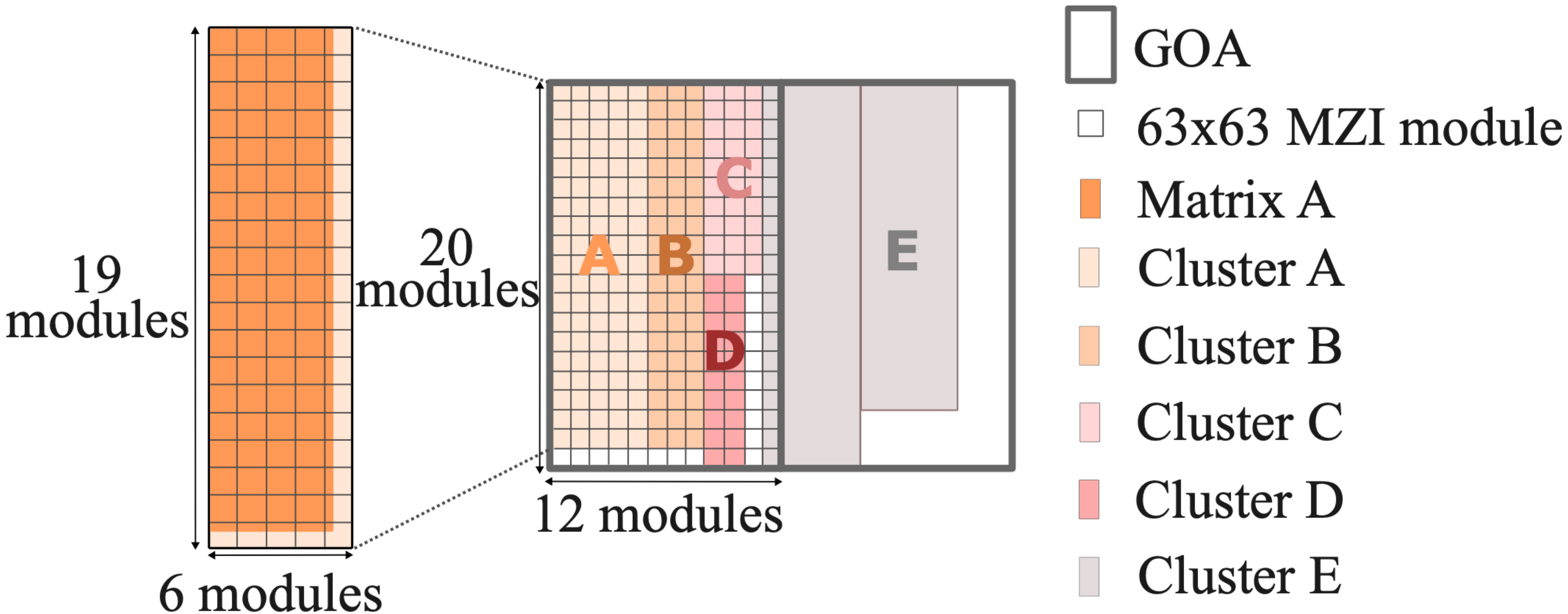
Represent a 4x4 matrix using an 8x8 GOA
(18+4)/28 MZIs are affected and unused

GOA Architecture



Higher utiliztiom efficiency and low mapping effort with independent MZI modules

Mapping NNs and Determining **GOA** Parameters

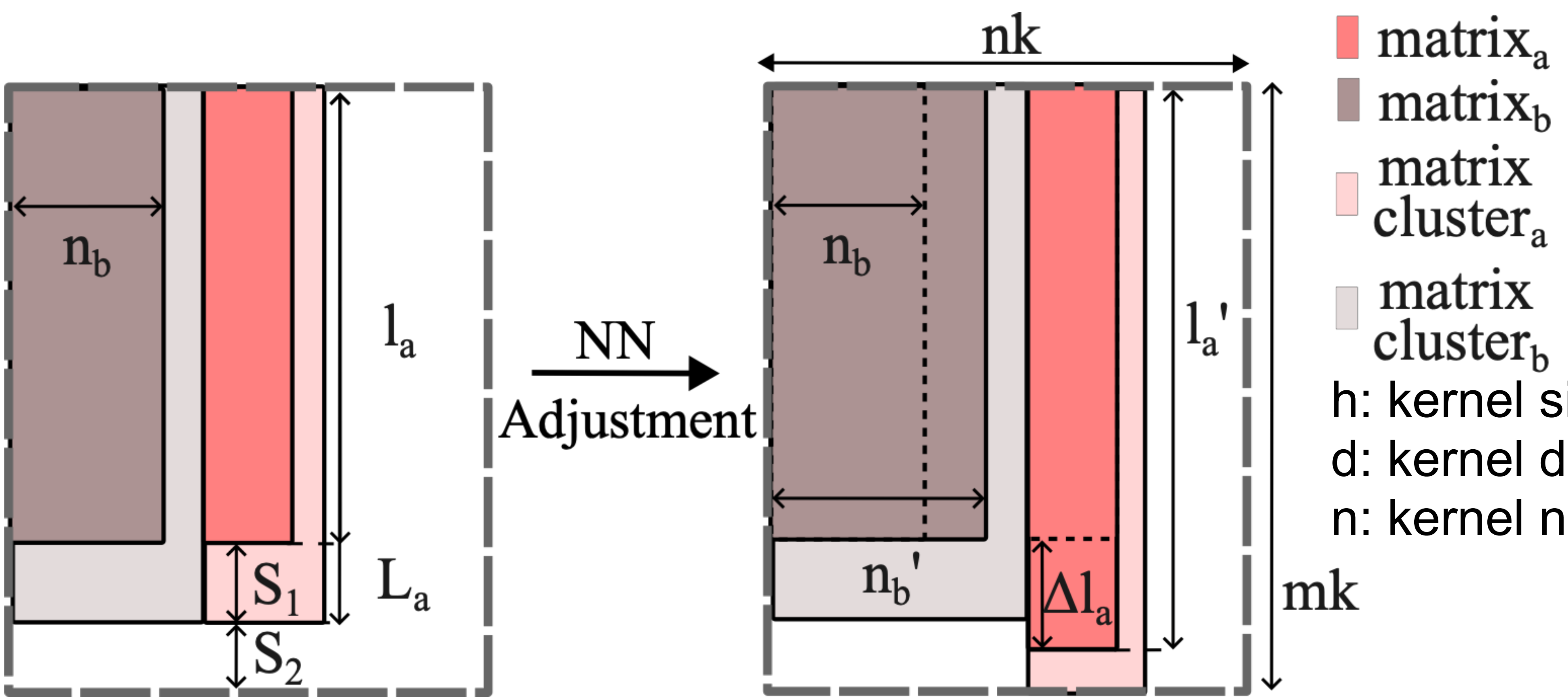


Genetic algorithm with metrics:

- Mapping cost - necessary mappings for one NN
- Area cost
- Power
- E/O conversions

MZIs, MRRs and peripheral devices

NN adjustment and Hardware-aware Training



l_a = h_a^2 x d_a
n_b = d_a

h: kernel size
d: kernel depth
n: kernel number

- Training

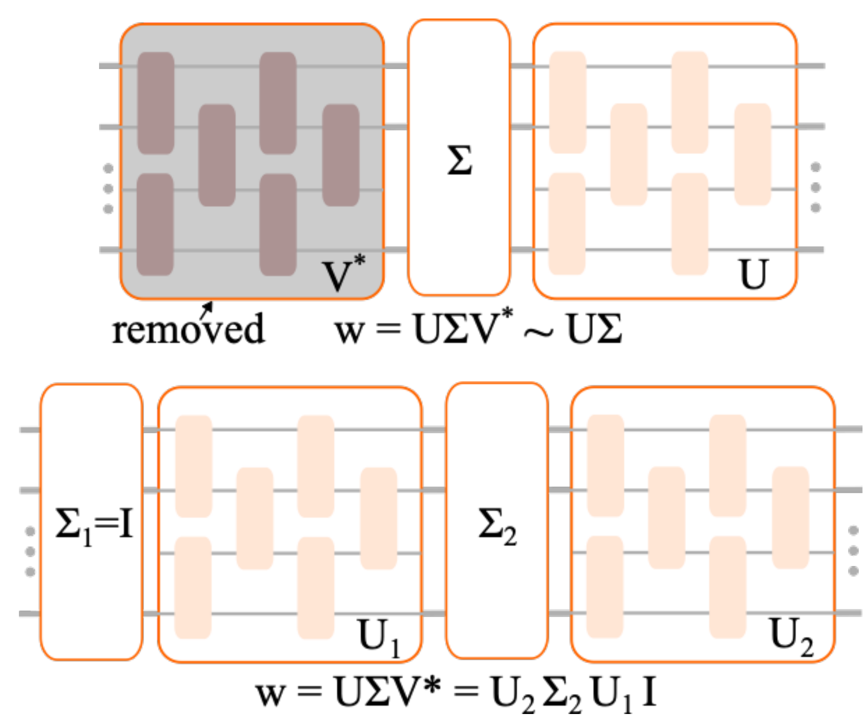
$$w' = U\Sigma$$

$$U = U_{svd}V^*_{svd}, \text{ where } w = U_{svd}\Sigma_{svd}V^*_{svd}$$

$$\Sigma = \begin{pmatrix} \sigma_{1,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{k,k} \end{pmatrix}.$$

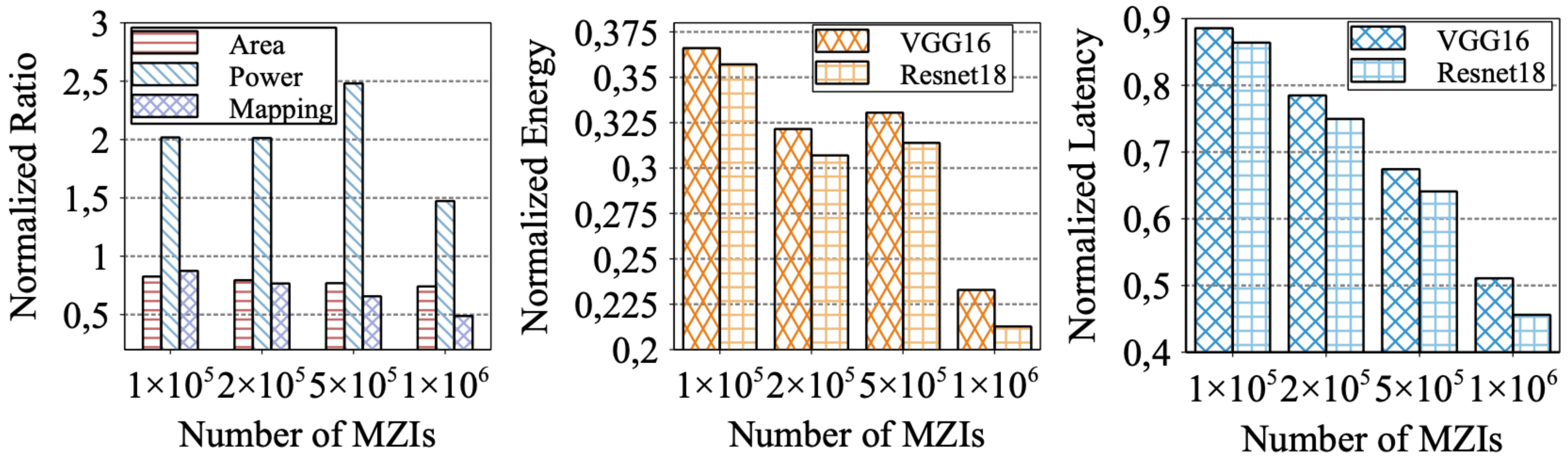
$$\sigma_{k,k} = \underset{\sigma}{\operatorname{argmin}}(\|w_k - \sigma_{k,k}U_k\|_2)$$

- Restoring by columns

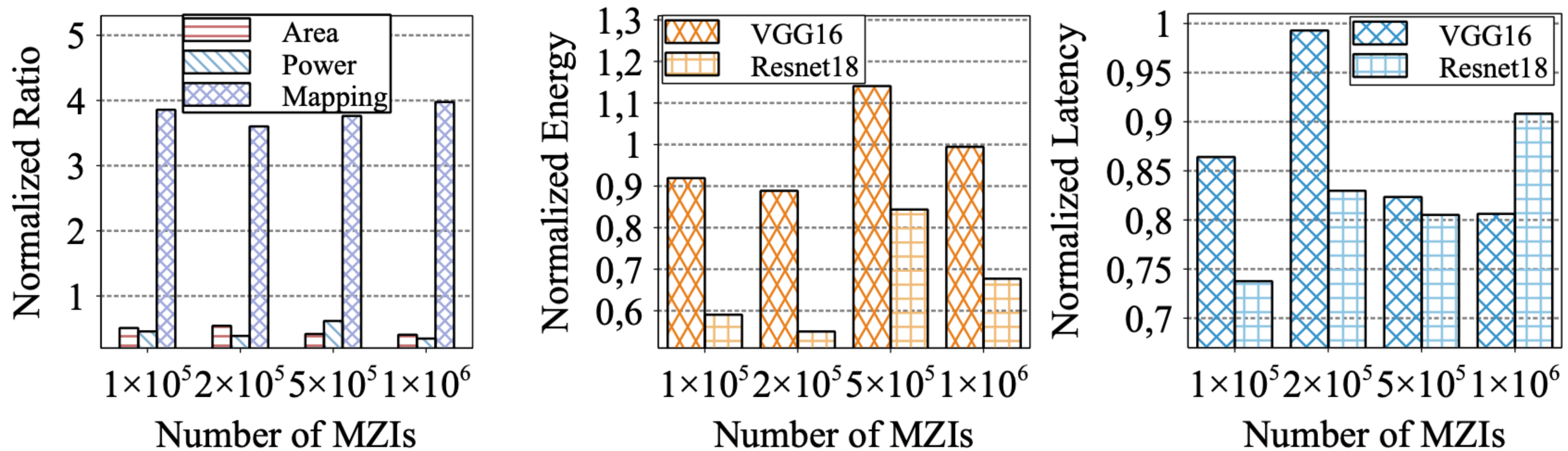


Experimental Results-Mapping/Energy/Latency Analysis

- Compared With SVD accelerators [6]



- Compared With Adept accelerators [12]

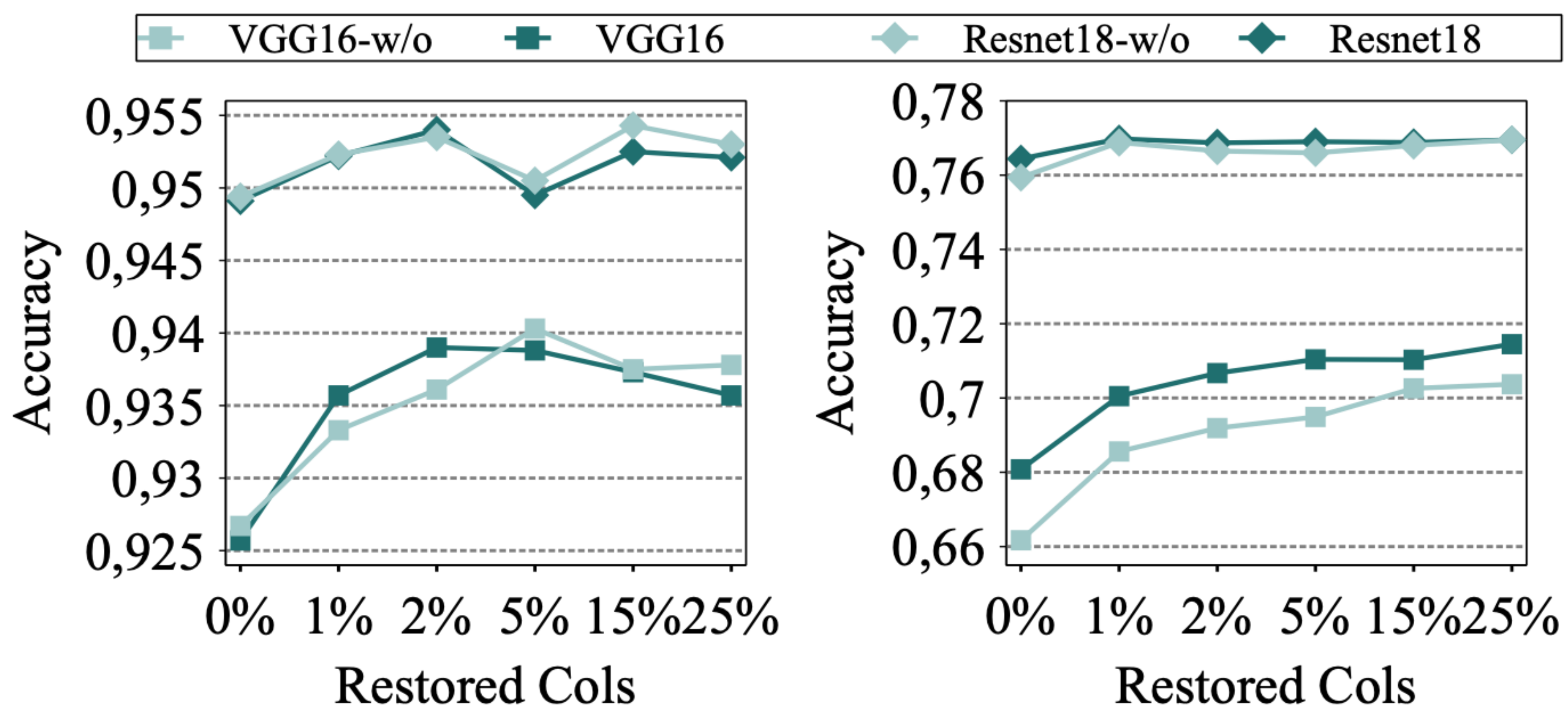


[6] YichenShen, NicholasCHarris, Skirlo, etal. Deep learning with coherent nanophotonic circuits. Nature photonics, 11(7):441–446, 2017.
[12] Jiaqi Gu, Hanqing Zhu, Chenghao Feng, et al. Adept: Automatic differentiable design of photonic tensor cores. In Design Automation Conference (DAC), 2022.

Experimental Results-Accuracy Analysis

Table 2: Results of the proposed framework. MZI number constraint: 20000, m, n, k = 6,3,44

Neural Networks Dataset	Performance Improvement			Accuracy				Restored Cols
	Mapping Reduction	Energy Reduction	Latency Reduction	Baseline	This Work w/o adjustment	This Work w/o restoration	This Work	
VGG16-Cifar10	21.87%	67.96%	21.85%	93.55%	92.57%	92.67%	93.57%	1%
VGG16-Cifar100	21.20%	67.71%	21.19%	70.16%	67.12%	68.35%	70.67%	2%
Resnet18-Cifar10	24.69%	69.13%	24.61%	94.93%	94.91%	94.94%	95.22%	1%
Resnet18-Cifar100	25.52%	69.47%	25.45%	75.79%	75.94%	76.44%	76.44%	0%



Cifar10 Cifar100

Conclusion

- To reduce mapping effort：
  - a GOA architecture of independent MZI modules is proposed
  - #params of GOA is determined by balancing the area cost, power, mapping cost, and E/O conversions
  - NN adjustments, hardware-aware training, restoration of weight matrices are performed to ensure accuracy
- Mapping efficiency improved up to 25.52%, energy saved up to 67%, latency saved up to 21%, compared to the SVD accelerator

Thank you!