ADEPT-Z: Zero-Shot Automated Circuit Topology Search for Pareto-Optimal Photonic Tensor Cores

Ziyang Jiang[1], Pingchuan Ma[1], Meng Zhang[2], Rena Huang[2], Jiaqi Gu[1]

[1]Arizona State University, [2]Rensselaer Polytechnic Institute

[1]School of Electrical, Computer and Energy Engineering

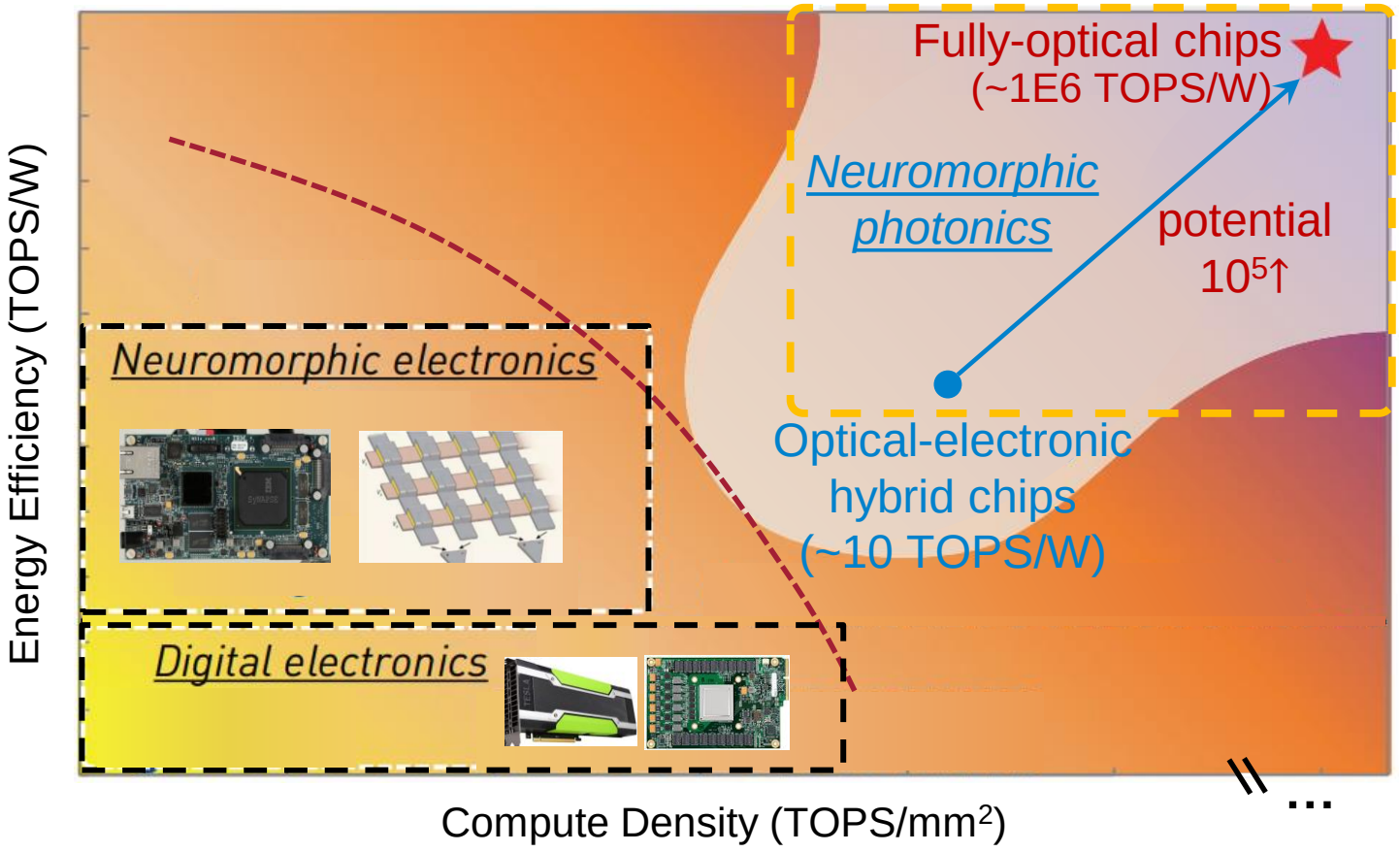
zjian124@asu.edu

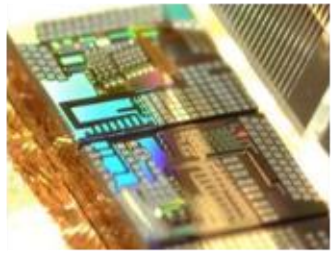
jiaqigu@asu.edu | *scopex-asu.github.io*

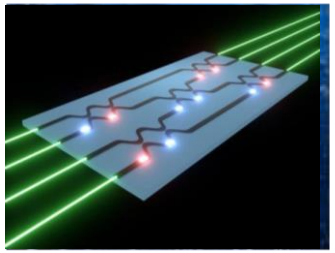
Photonic ML Accelerators

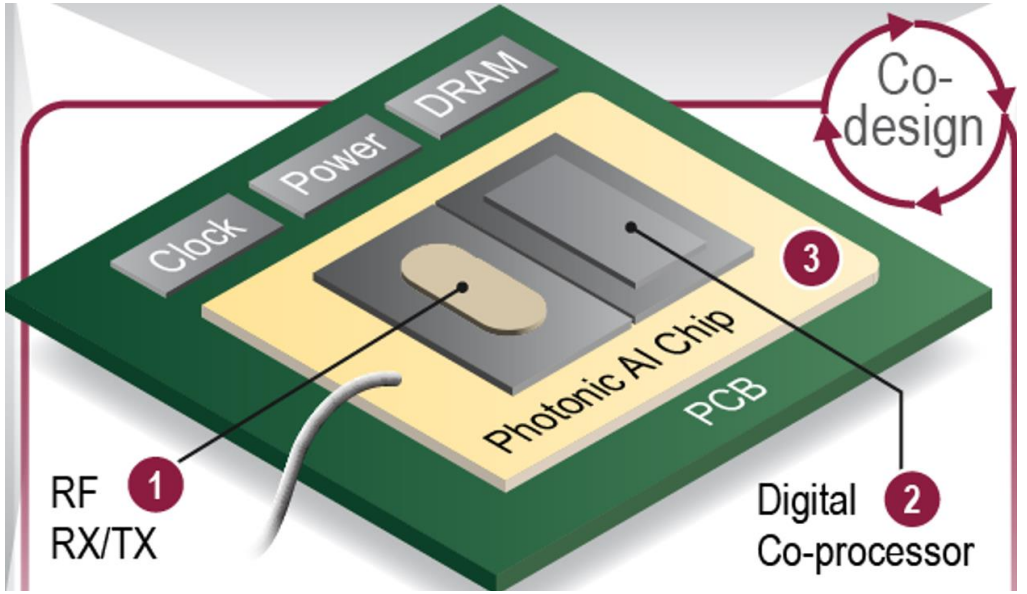
- Evolve from electronics to *integrated photonics*

Speed-of-light, Massive parallelism, low-power







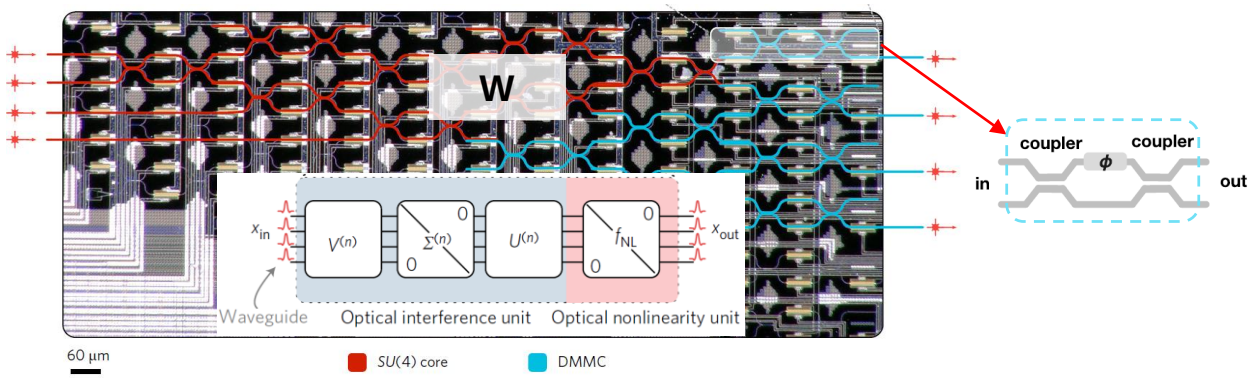


Source: Mitchell A. Nahmias, Bhavin J. Shastri, Alexander N. Tait, Thomas Ferreira de Lima and Paul R. Prucnal, "Neuromorphic photonics," Optics & Photonics News, Jan 2018.

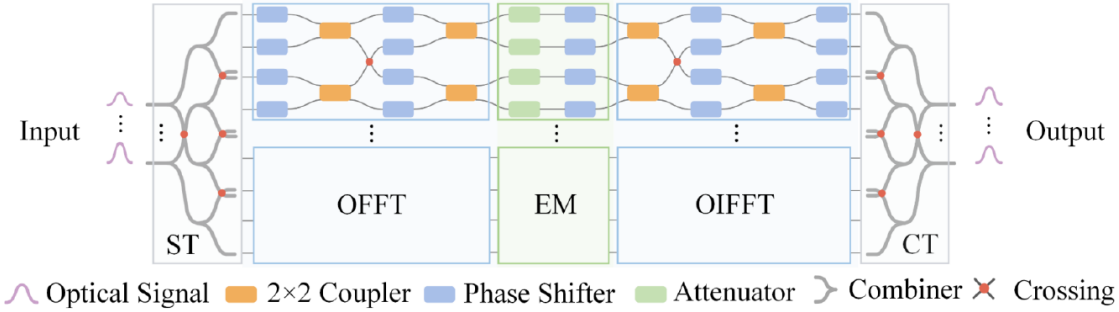
Copyright © 2022 Arizona Board of Regents

PTC: Photonic Tensor Core

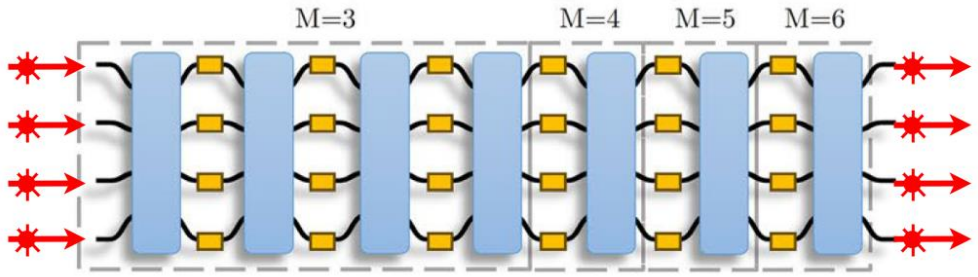
Coherent PTCs: Complex-valued Matrix Multiplication



SVD-based ONNs with MZIs [Nature'17, Shen+]



Butterfly-style PTC [ASP-DAC'20, Gu+]



PTC with MMI Devices [Sci. Rep.'24, Zelaya+]

Modular Coherent PTC Design

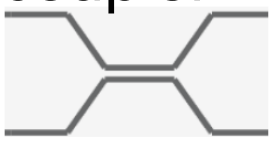
Repeated block structures

- Phase shifter arrays ($\Phi's$)
- Coupler arrays ($\mathcal{T}'s$)
- Crossing arrays ($\mathcal{P}'s$)

Matrix: $\mathbf{U} = \mathcal{P}_B\mathcal{T}_B\Phi_B \ldots \mathcal{P}_2\mathcal{T}_2\Phi_2\mathcal{P}_1\mathcal{T}_1\Phi_1$

Basic Optical Components

DC coupler



$$\begin{pmatrix} t & \sqrt{1-t^2}\,j \\ \sqrt{1-t^2}\,j & t \end{pmatrix}$$

CR crossing

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

PS phase shifter

$$y = e^{-j\phi}x$$

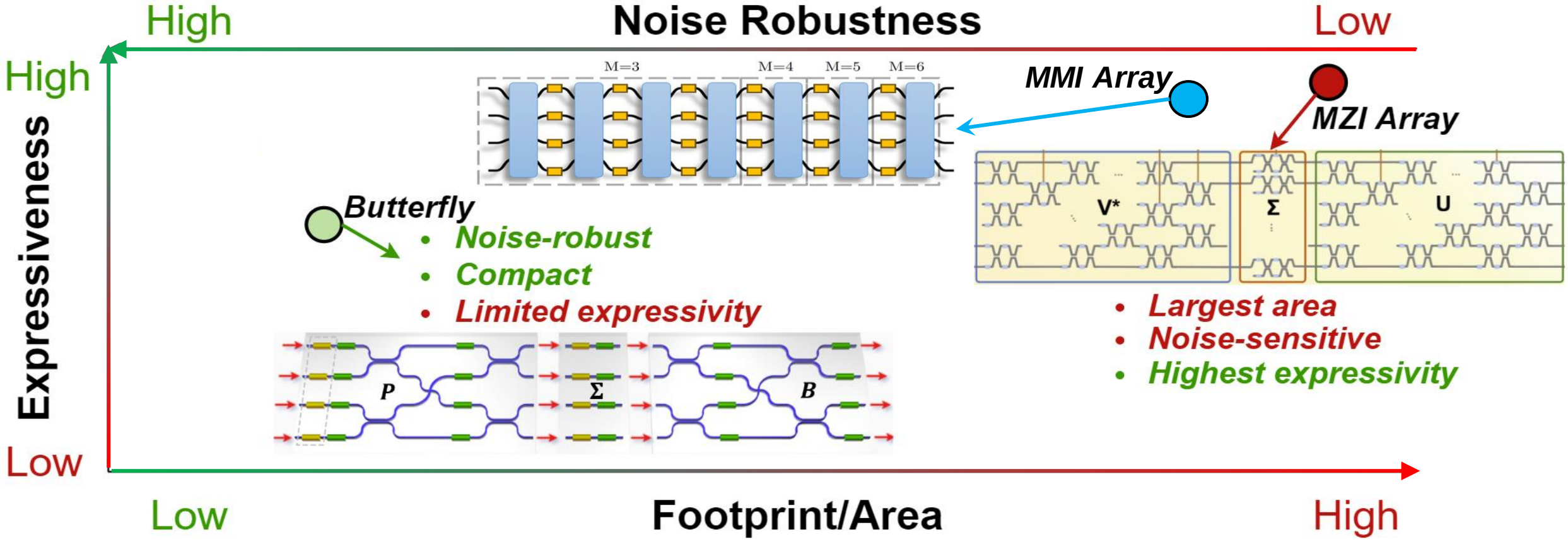
Copyright © 2022 Arizona Board of Regents

Manual PTC Design → Auto-Searched Design?

DA for Photonics

Manual Design → Automated PTC Design

Can we automatically & efficiently find Pareto-optimal photonic circuits designs?



Copyright © 2022 Arizona Board of Regents

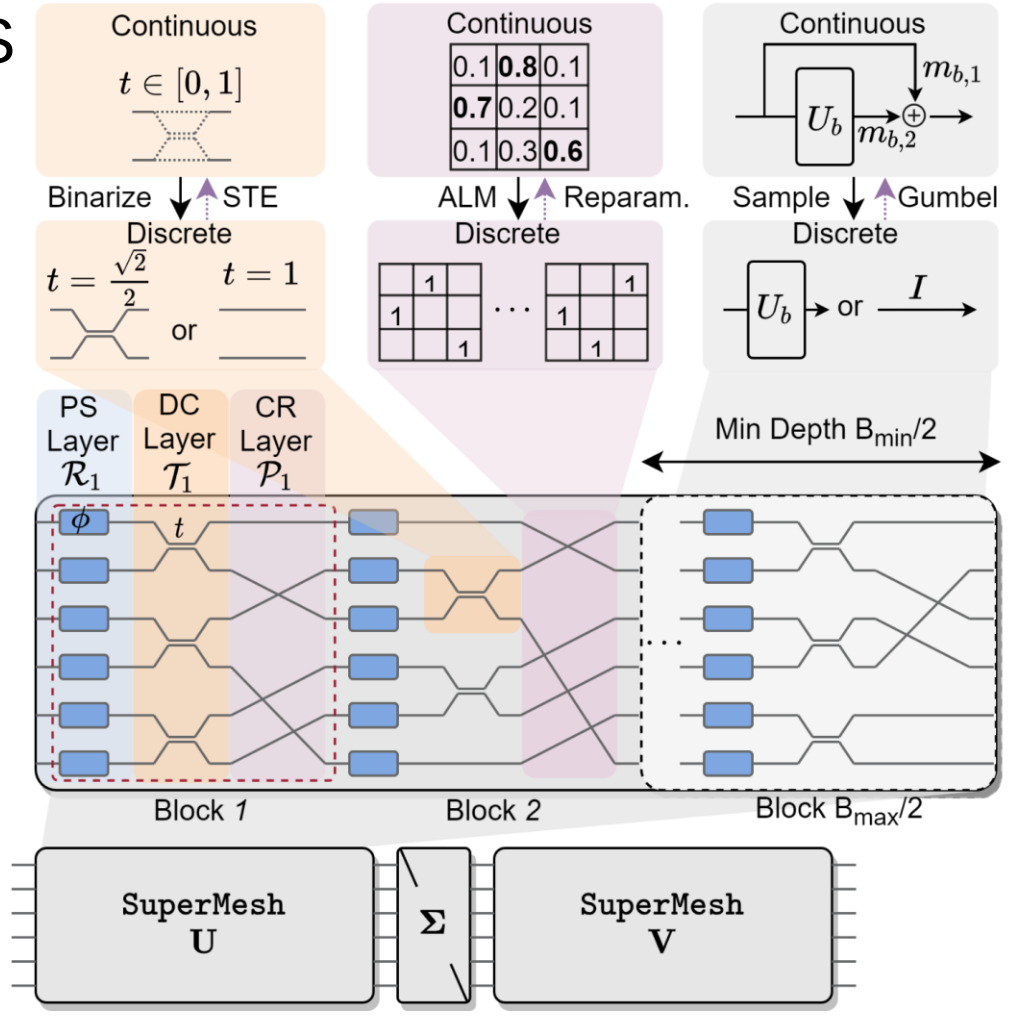
Prior Work on Auto PTC Design

- Relax as *continuous* optimization
- Gradient descent to search (similar to DARTS in *one-shot* neural architecture search)

Challenges in ***Differentiable*** ADEPT

- **Limited design space:** fixed slots for 2x2 coupler, *cannot handle multi-port couplers*
- **Optimization difficulty:** hard to handle multiple non-differentiable objectives / constraints Relaxation is *inaccurate and hard to converge*.
- **High search cost:** need to re-train for each search specification with *high training cost*.

ADEPT [DAC'22]
Differentiably auto-searched PTC

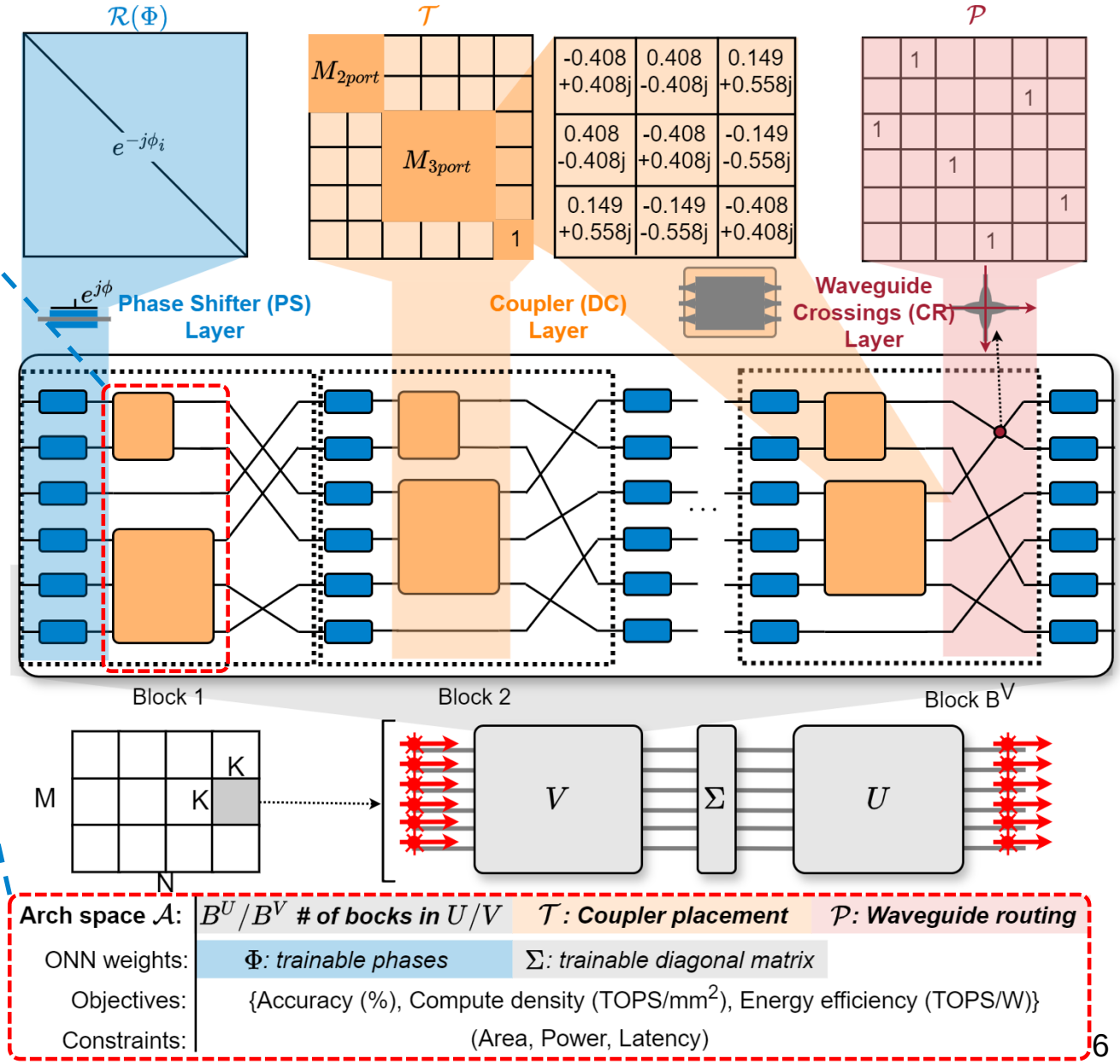


[ADEPT, Gu+, DAC'22]

Copyright © 2022 Arizona Board of Regents

Proposed Zero-Shot PTC Search: ADEPT-Z

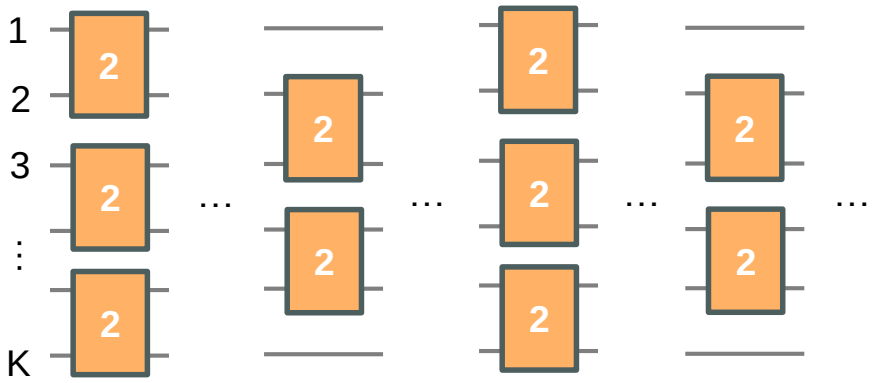
- **Extended** search space
  - Arbitrary placement of multi-port couplers
- **Support multiple** objectives / constraints
  - Accurate hardware cost modeling without surrogate model or relaxation
  - e.g., power, area, latency,
- **Efficient zero-shot** search
  - Training-free during search
  - Efficient accuracy proxy



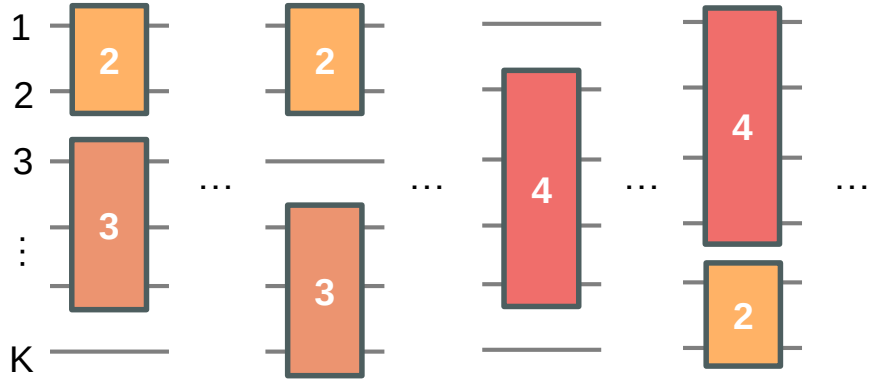
Copyright © 2022 Arizona Board of Regents

Extend Architecture Design Space ($\alpha \in \mathcal{A}$)

- ADEPT [DAC'22]
  - Only support **2-input** couplers with **fixed** placements
  - **Design space**: $O\left(\left(\frac{K}{2} \cdot K!\right)^{B_{max}}\right)$



- **ADEPT-Z**
  - Support **multi-port** couplers with **arbitrary** placements
  - **Design space↑**: $O\left(\left(2^{K-1} \cdot K!\right)^{B_{max}}\right)$
  - Dense signal interaction → reduce circuit depth



This exponential, discrete design space cannot be searched by gradient descent
ADEPT-Z uses evolutionary algorithm to explore this space

Copyright © 2022 Arizona Board of Regents

Overview of Zero-Shot PTC Topology Search

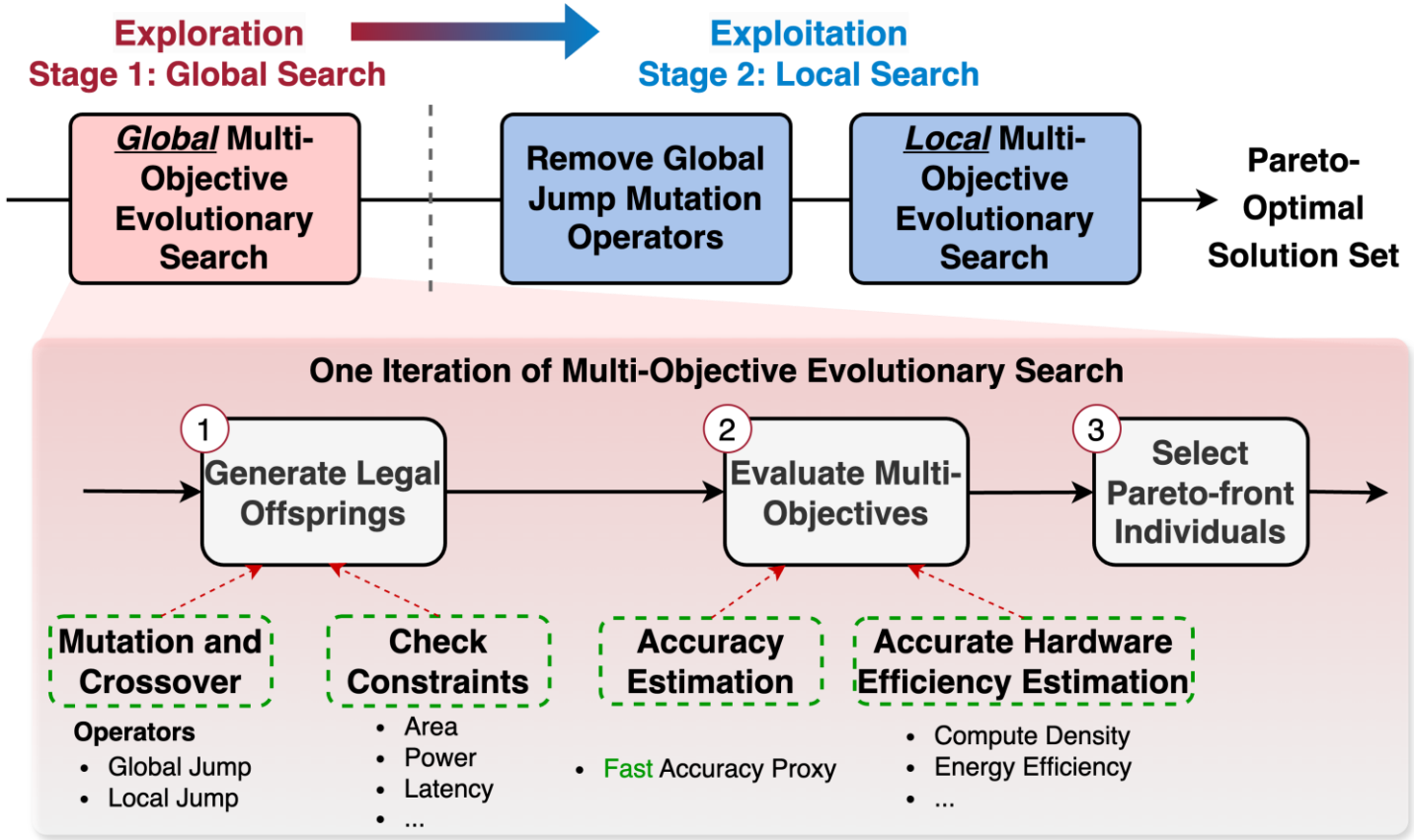
- Formulation

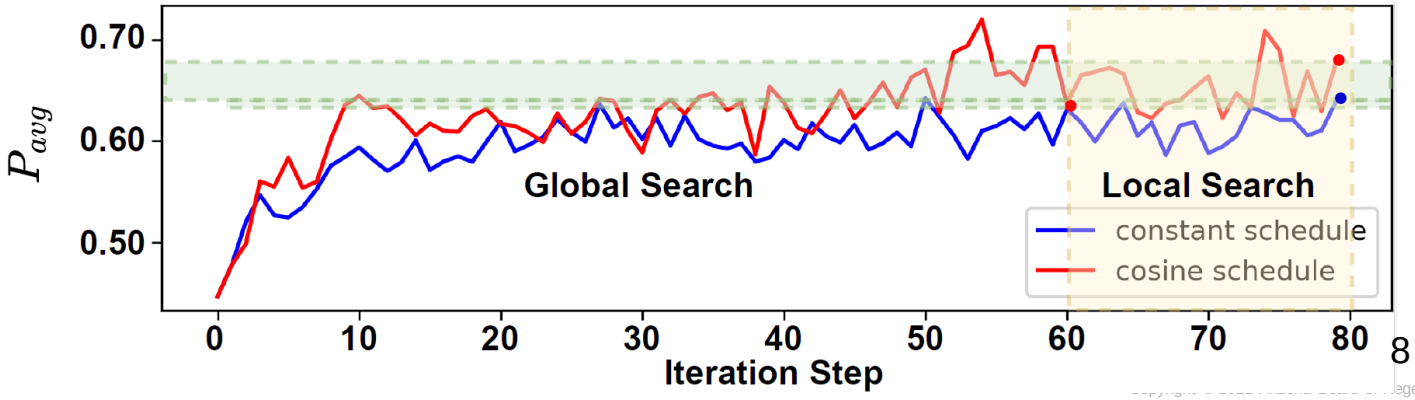
$$\max_{\alpha \in \mathcal{A}} (\text{Efficiency}, \text{Density}, \text{Accuracy})$$
$$\text{s.t.,}\ \ \text{Pmin} < \text{Power}(\alpha) < \text{Pmax}$$
$$\text{Amin} < \text{Area}(\alpha) < \text{Amax}$$
$$\tau_{min} < \text{Latency}(\alpha) < \tau_{max}$$

- Multi-objective: NSGA-II Pareto front search
- Accuracy proxy: fast training-free acc. estimation
- Hardware constraints: posed by population filtering

One-shot Search: 6-10 hours to get a single solution

Zero-shot Search: 3 hours to get multiple pareto-optimal solutions



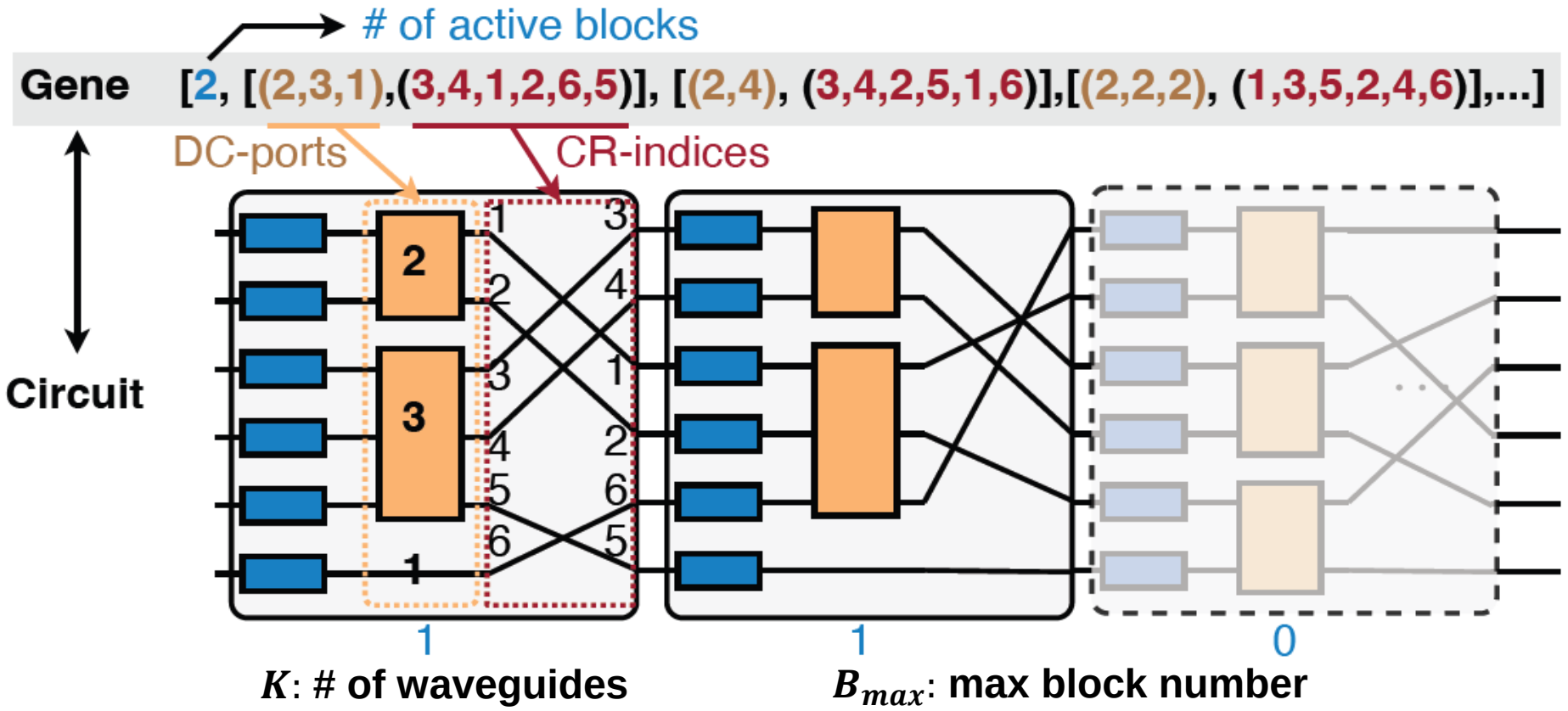


Multi-Objective Evolution: Gene Representation

- Compact *Gene-to-Circuit* mapping
  - #activeBlk-DC-CR-DC-CR-….
  - DC: port count list, e.g., [2,3,1], sum to K
  - CR: output port indices, e.g., [3,4,1,2,6,5]

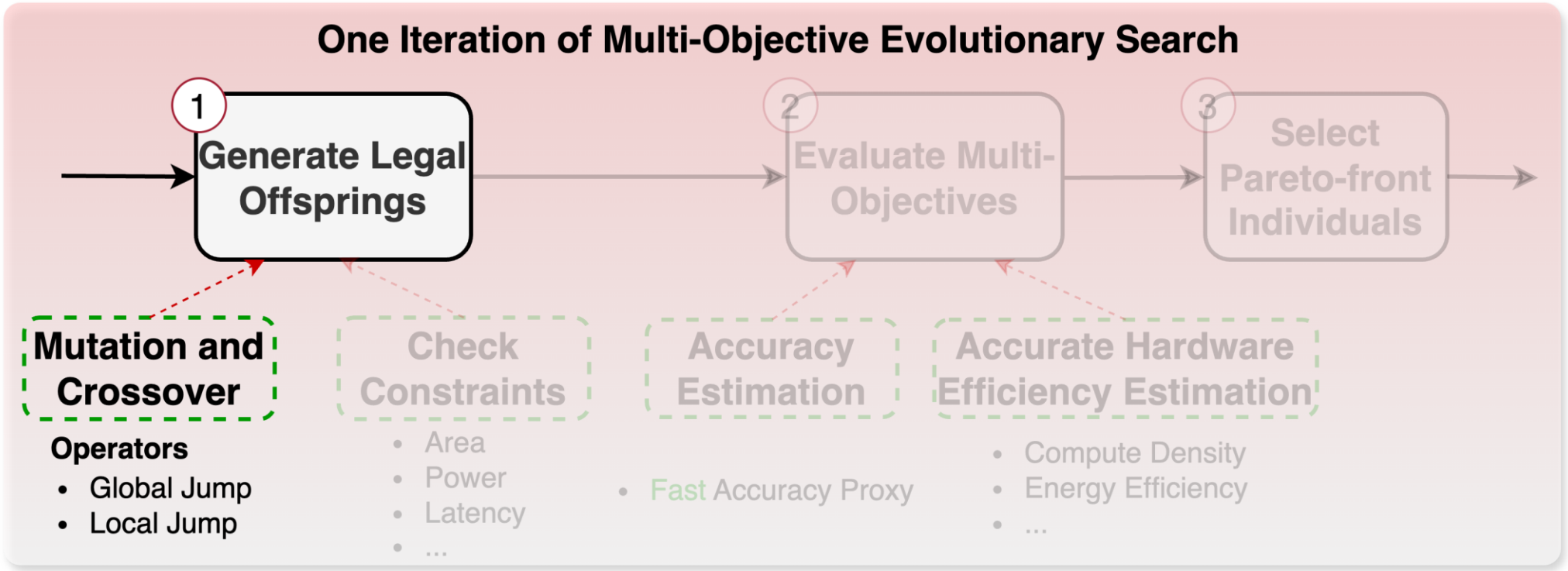
Compact representation of multi-port couplers

Easy to manipulate in evolution



Copyright © 2022 Arizona Board of Regents

Proposed Zero-Shot PTC Topology Search



Copyright © 2022 Arizona Board of Regents

Exploration in Large Design Space: Mutation

- Customized mutation operators: *legal move* + *sufficient change* + *good stability*
- Step 1 Global Search: Global Jump + Local Jump
- Step 2 Local Search: Local Jump only

Balance between
Exploration and **Exploitation**

Local Jump

Global Jump

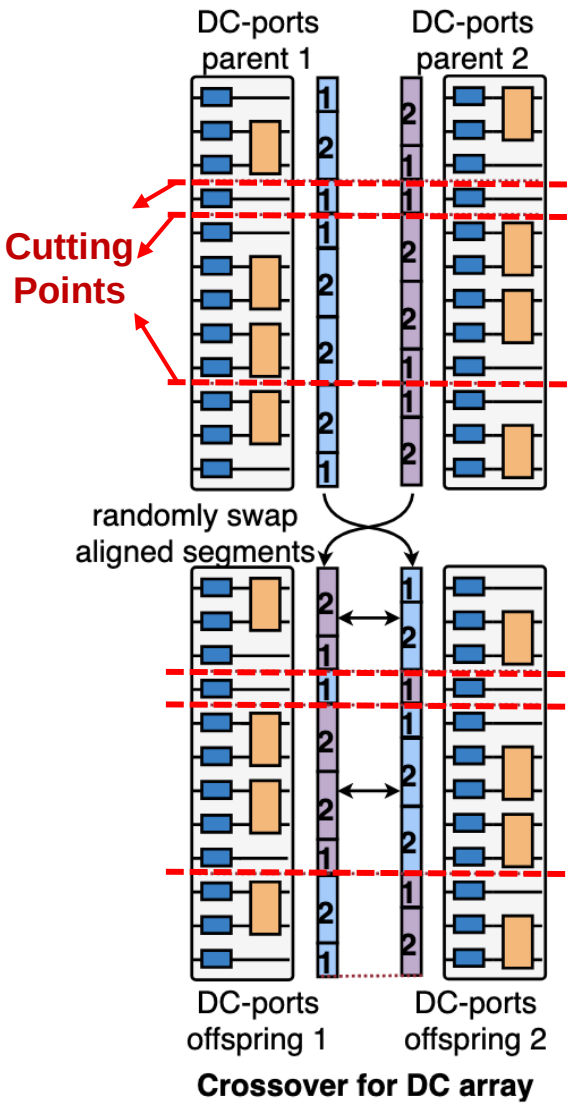
Type	Mutation Ops	Description
DC	R2A1	Remove 2 and add 1 DC. E.g., [2,4,2] → [1,1,1,1,1,1,2] → [1,4,1,2]
	A2R1	Add 2 and remove 1 DC. E.g., [1,1,2,1,1,1,1] → [1,1,2,2,2] → [1,1,2,1,1,2]
	Move	Move one DC to another position. E.g., [1,2,2,2,1] → [1,2,2,1,2]
	RS	Resample a DC array
CR	AddCR	Add a random number of CR. E.g., Add 2 CRs: [0,1,2,3] → [0,3,1,2]
	ReduceCR	Reduce a random number of CR. E.g., Reduce 1 CR: [0,2,1,3] → [0,1,2,3]
Block	AddBlock	Copy a random number of blocks from the front of gene, add to the end
	ReduceBlock	Remove a random number of blocks from the end of gene.

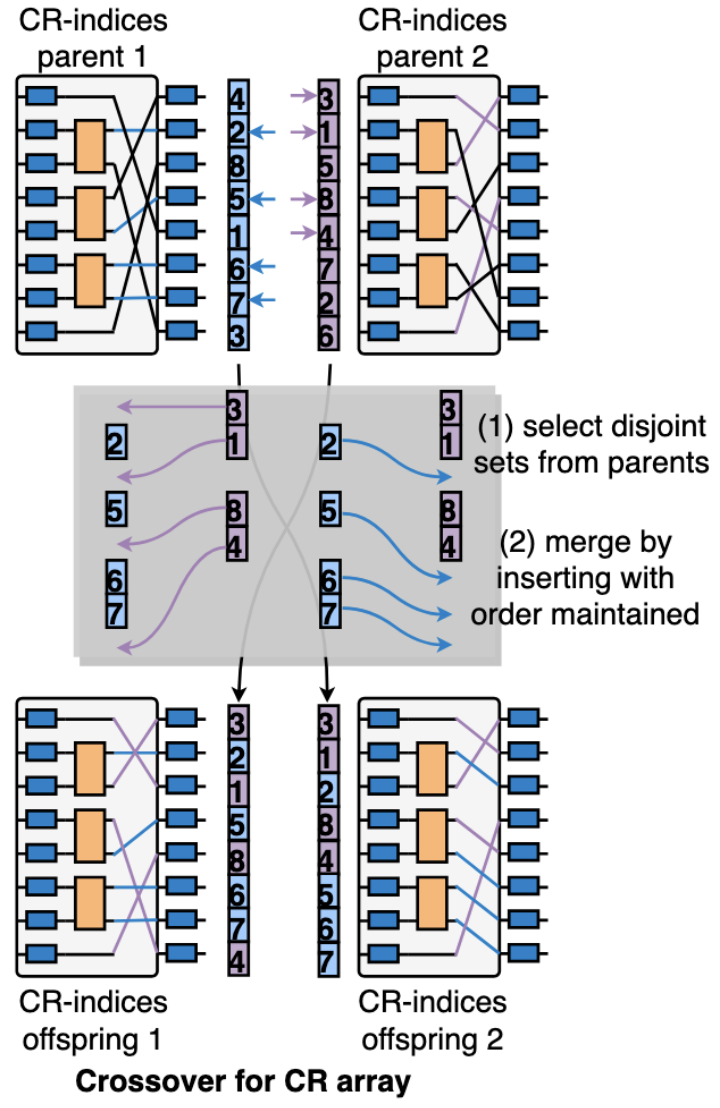
Copyright © 2022 Arizona Board of Regents

Exploitation in Large Design Space: Crossover

- Guarantee **gene legality**
- Maintain **good property of parents (interpolation)**

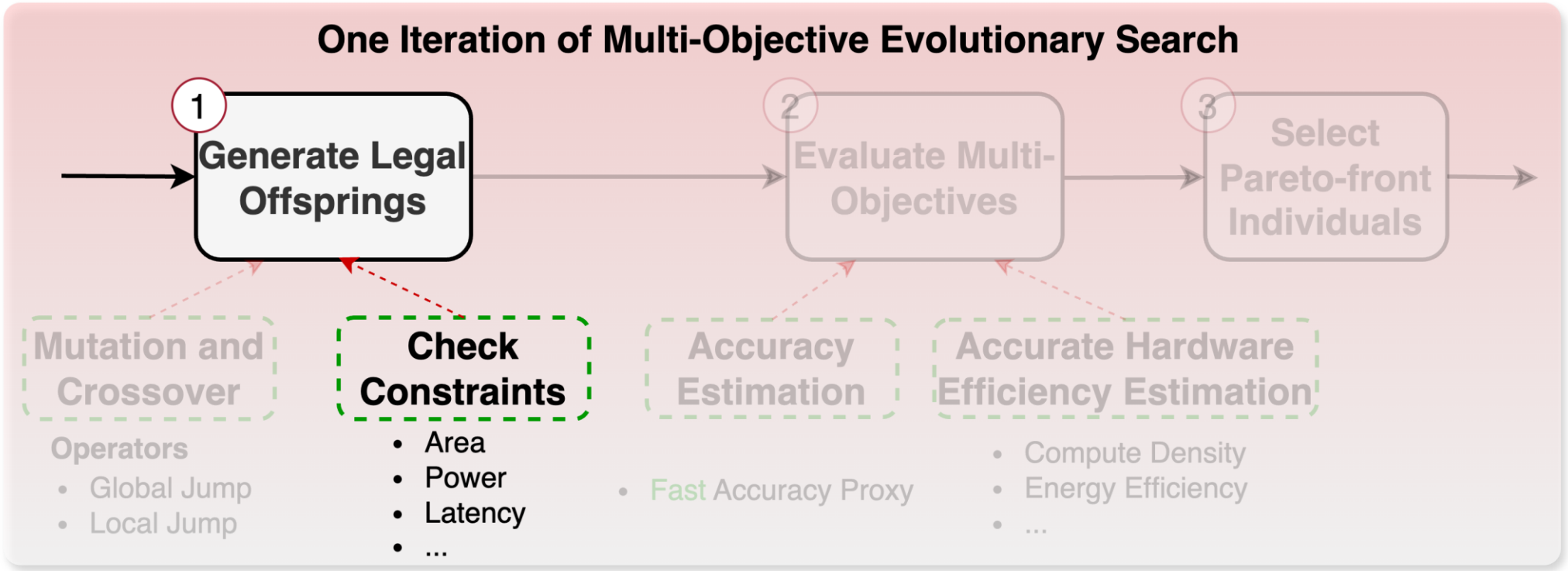
- DC crossover
  - Genes → Boarder-Aligned Segments
  - Swap aligned segments randomly
- CR crossover
  - Select complementary subset of indices from parents
  - Cross-insert and preserve the relative order in parent genes
- Block crossover
  - Swap two active blocks randomly





Copyright © 2022 Arizona Board of Regents

Proposed Zero-Shot PTC Topology Search



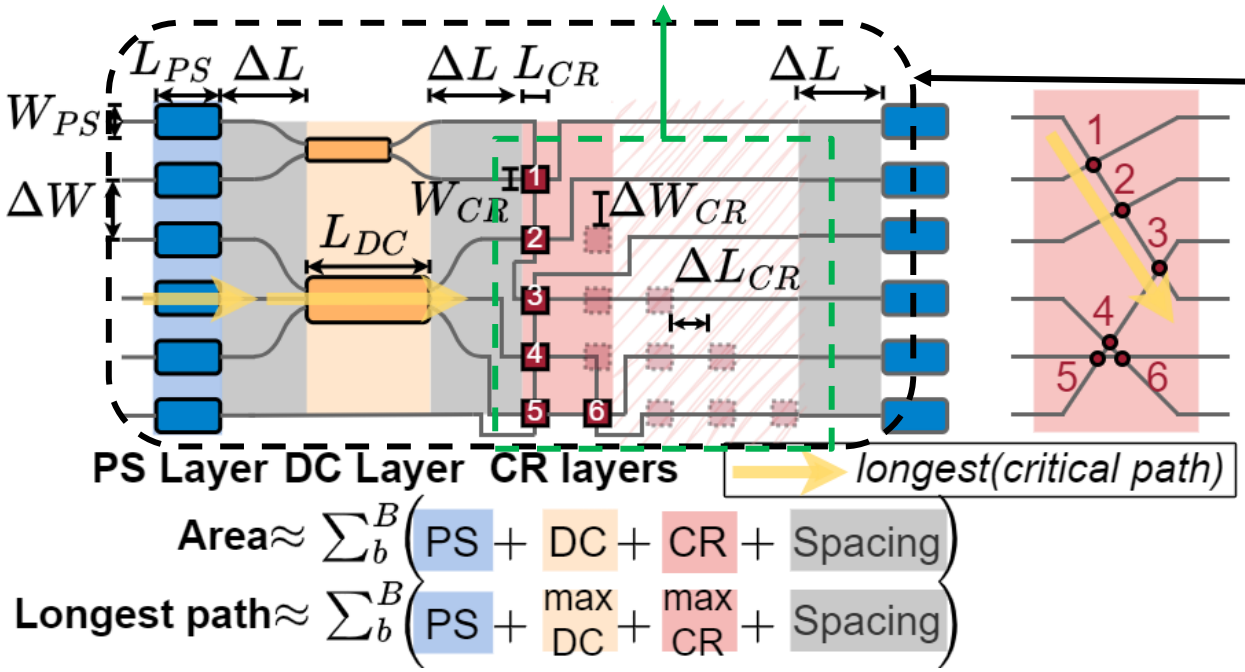
Copyright © 2022 Arizona Board of Regents

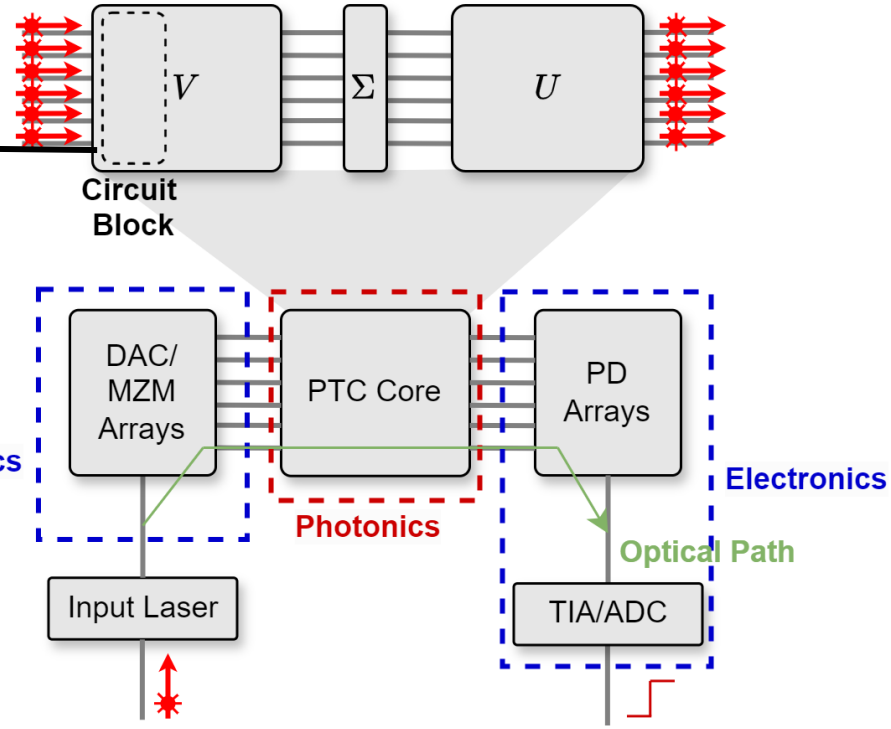
Explicit Hardware Constraints in the Loop

- **Layout-aware** area estimation
  - $A(\alpha) = A_U(\alpha) + A_V(\alpha) + A_\Sigma + K(A_{TIA} + A_{PD} + A_{MZM} + A_{DAC} + A_{ADC})$
    - Photonics Area
    - Electronics Area
  - Consider **actual chip layout**, practical **spacing**

More **accurate** estimation than summing optical device **footprints**

Predefined **triangular** layout for crossing arrays





Copyright © 2022 Arizona Board of Regents

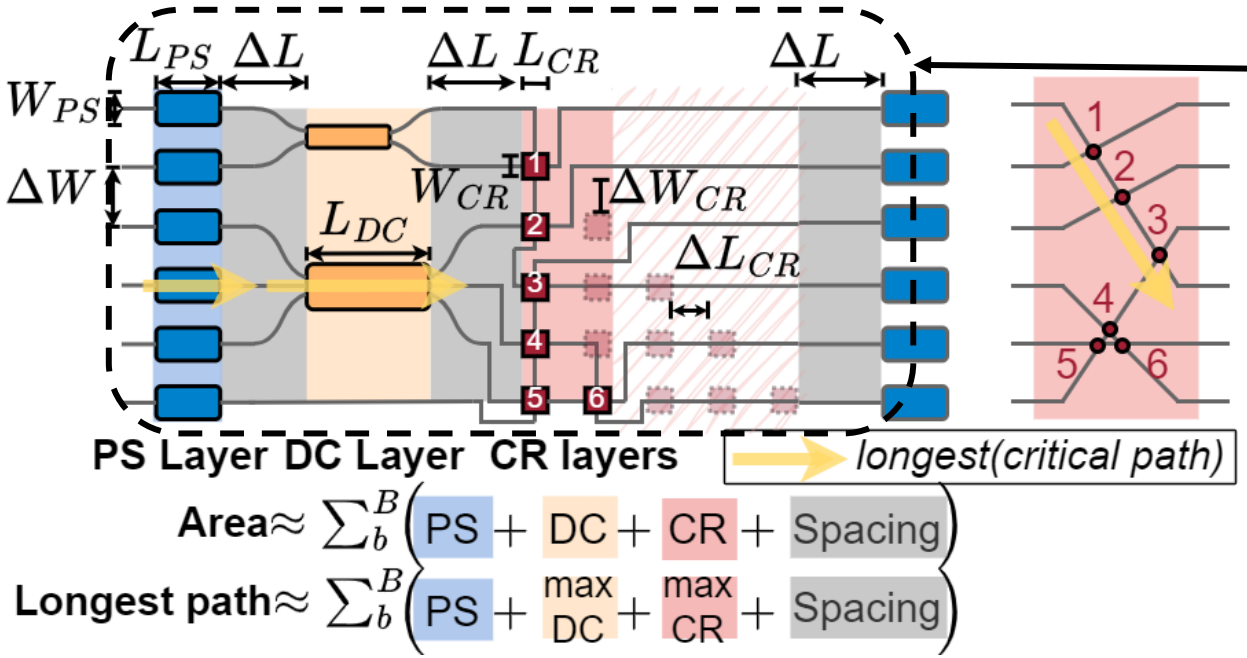
Explicit Hardware Constraints in the Loop

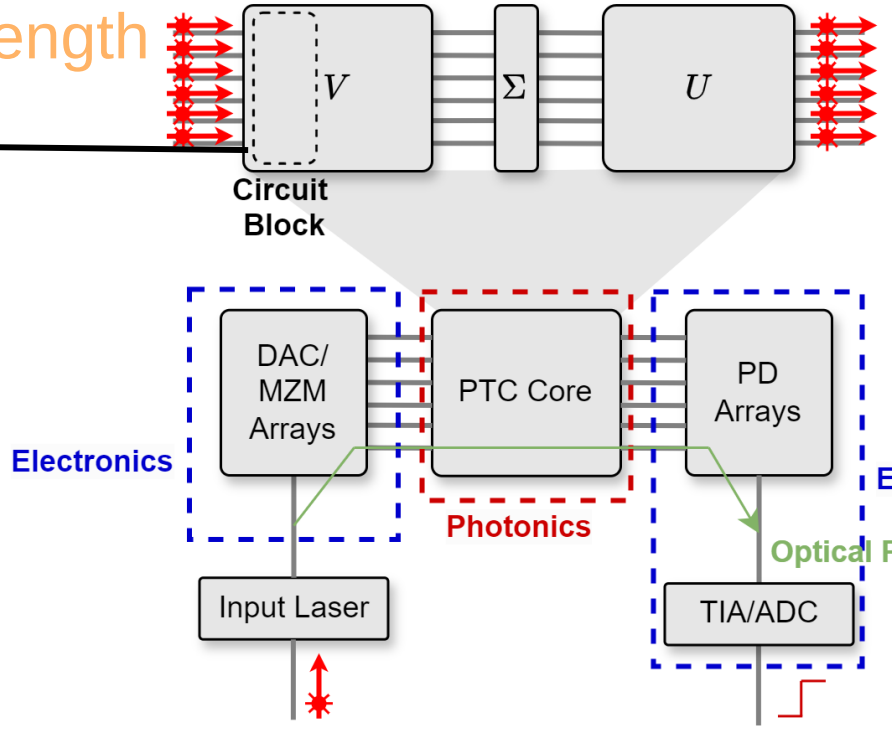
- **Total electrical/optical** power estimation
  - $P(\alpha) = P_{laser} + K \cdot (P_{MZM} + P_{DAC} + P_{ADC} + P_{TIA} + P_{PD})$

Laser Power Electrical Circuit Power

- **Critical path-aware** latency estimation
  - $\tau(\alpha) = \max(f^{-1}, n_g L_{path}/c_0 + \tau_{DAC} + \tau_{PD})$

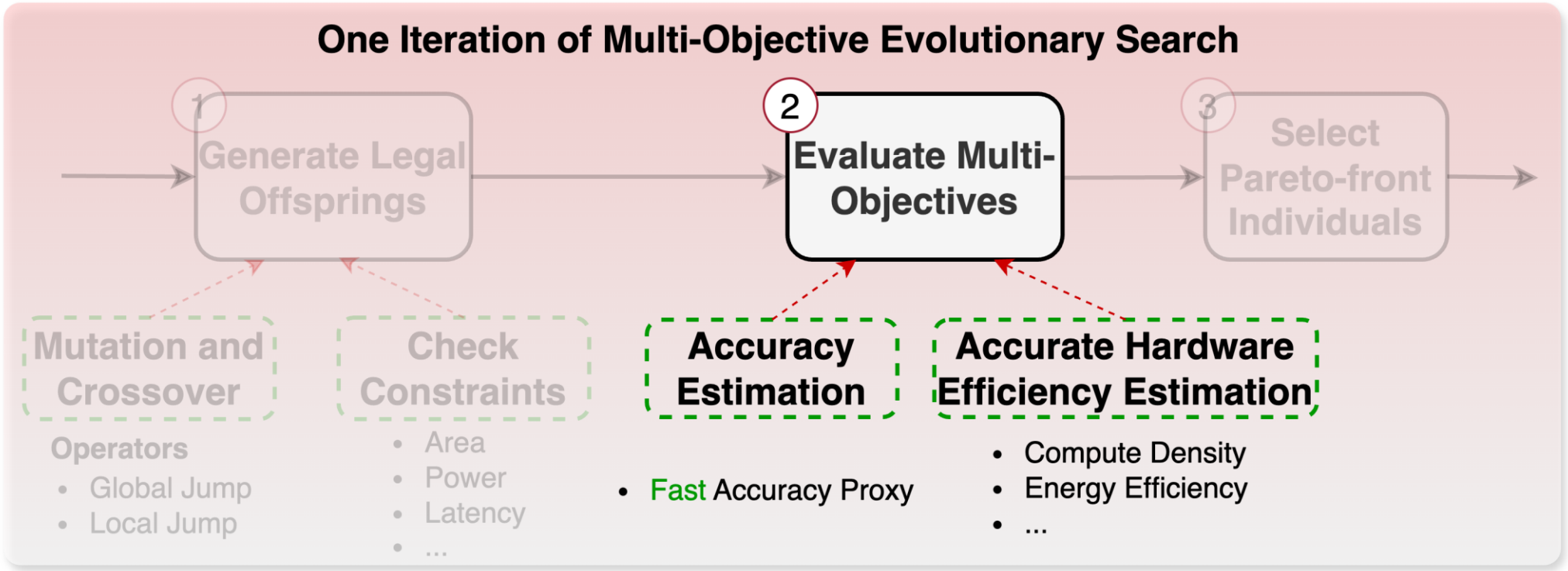
Clock rate **Longest** Optical Path Length **Total** Optical Path Delay





Copyright © 2022 Arizona Board of Regents

Proposed Zero-Shot PTC Topology Search



Copyright © 2022 Arizona Board of Regents

Hardware Cost & Training-Free Acc. Estimation

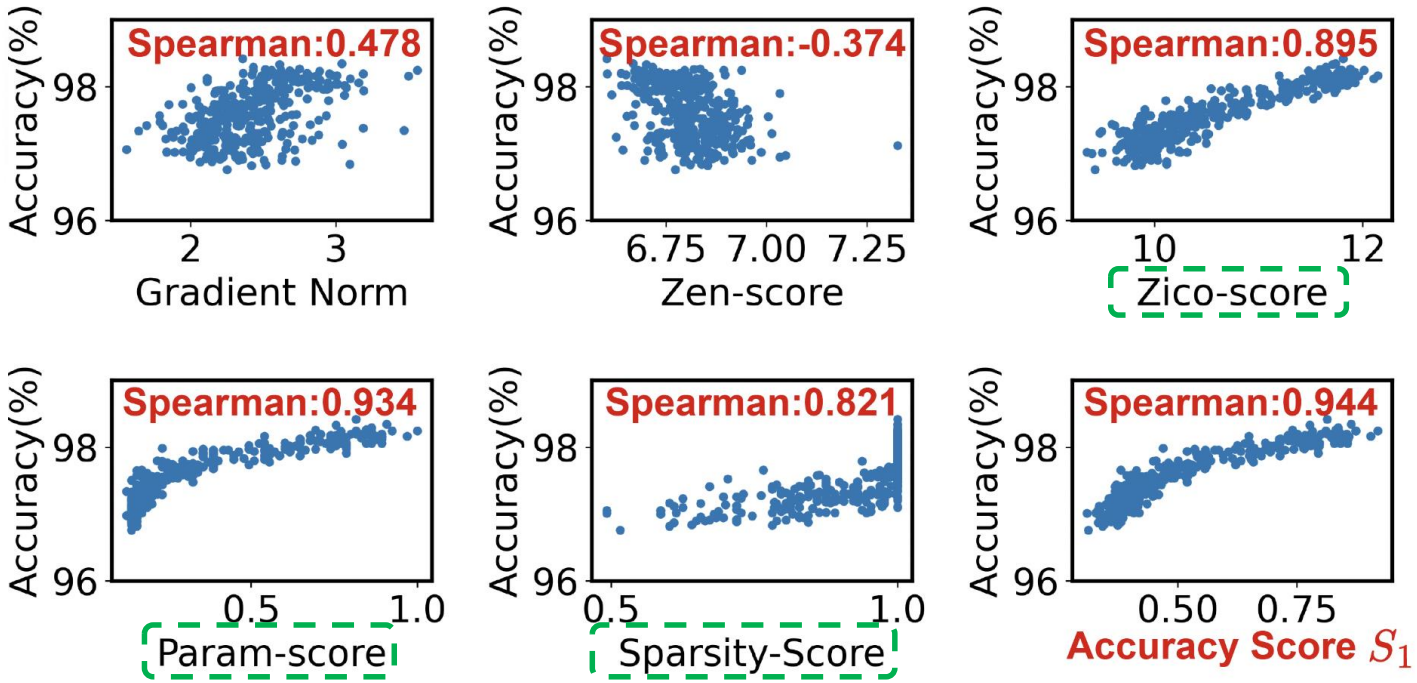
- Energy efficiency (TOPS): $EE(\alpha) = 2K^2/(\text{Power}(\alpha) * \text{Latency}(\alpha))$
- Compute density (TOPS/mm²): $CD(\alpha) = 2K^2/(\text{Area}(\alpha) * \text{Latency}(\alpha))$

- Accuracy: Find 3 easy-to-compute accuracy indicators
  - High rank correlation with true test accuracy (spearman >0.94)
  - Comprehensive accuracy Score S_1 w/ regressed coefficients.

Training v.s. Accuracy Estimation
>30min → **10 Seconds**

- ZiCo [Li, ICLR'23]: grad property (trainability, generalization)
- #Params in circuit/Sparsity of matrix: (expressivity)

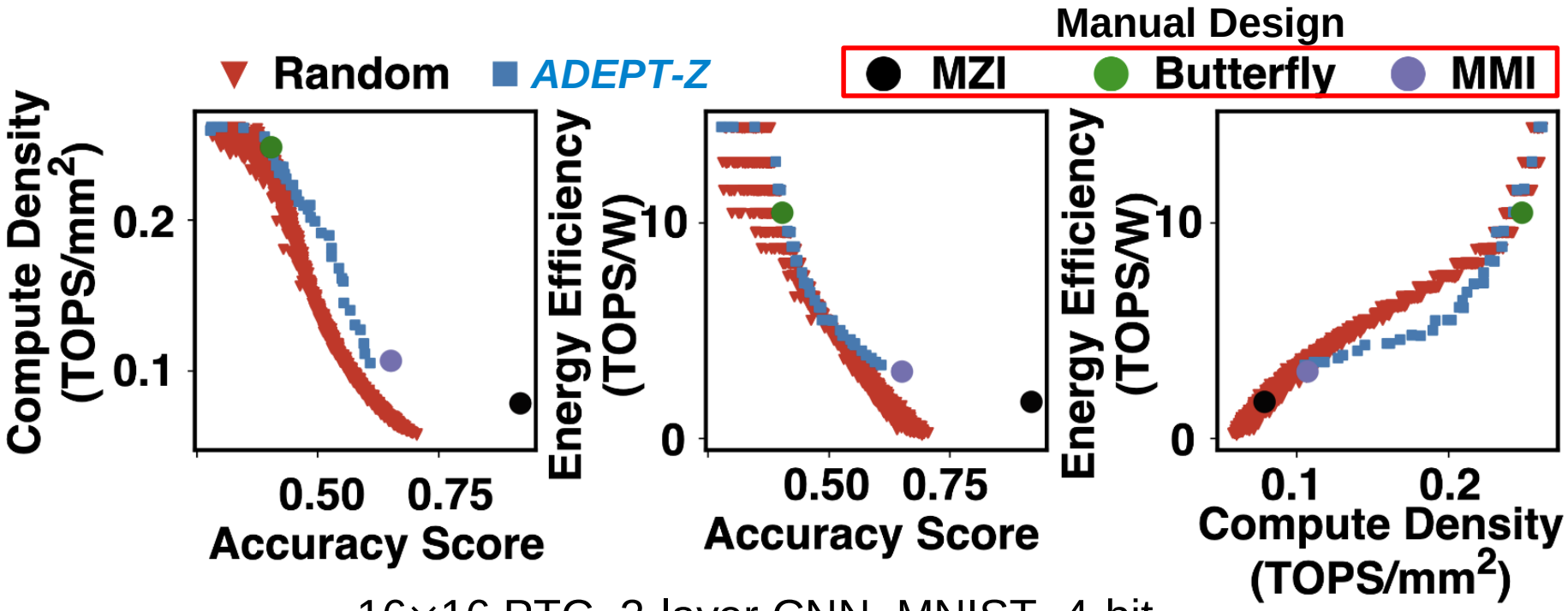
$$\mathcal{S}(\alpha) = 0.015 \cdot S_{\text{Zico}}(\alpha) + 0.561 \cdot S_{\text{Param}}(\alpha) + 0.175 \cdot S_{\text{Sparsity}}(\alpha)$$



Copyright © 2022 Arizona Board of Regents

Auto-Searched Pareto-Optimal PTCs on GF PDK

- Search 16×16 PTC (2/8-port DC) on GlobalFoundries PDK
- Pareto front covers all three manual designs
- Better than **Random** search
- **One-time** search, **40** Pareto-optimal solutions within **2.7** hours
  - 100× faster than ADEPT (6-10 hours per one-shot differentiable search)



16×16 PTC, 2-layer CNN, MNIST, 4-bit

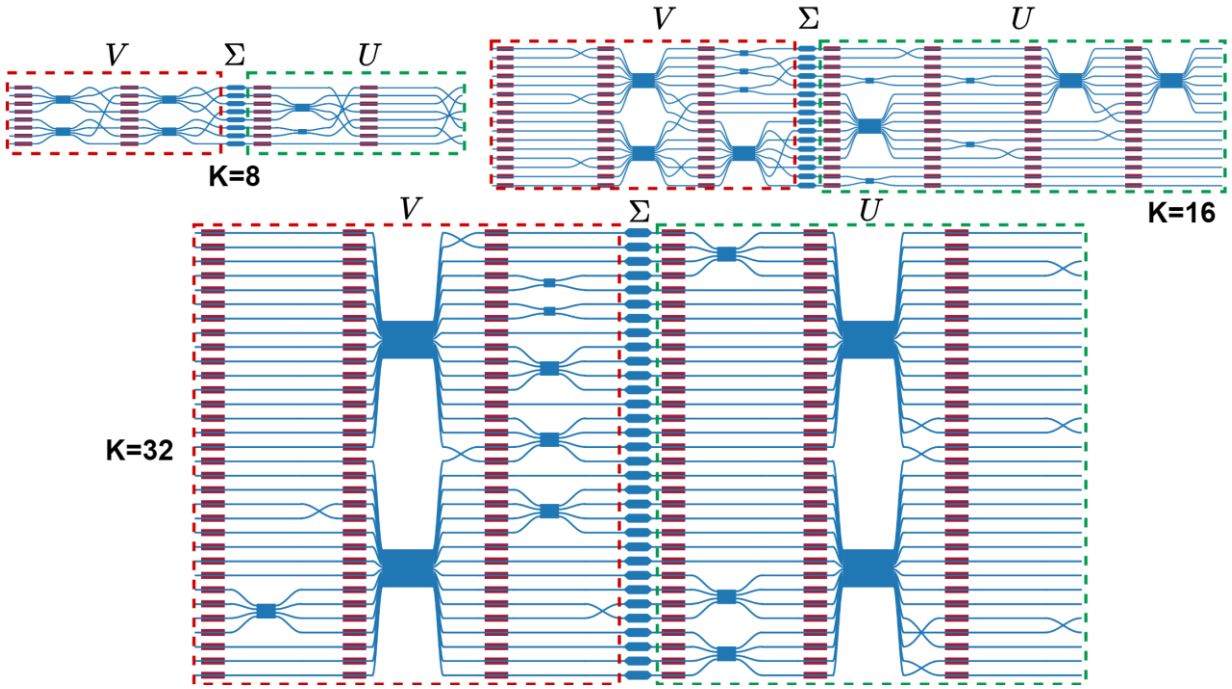
Copyright © 2022 Arizona Board of Regents

Auto-Searched Pareto-Optimal PTCs on GF PDK

- Pareto-optimal PTC ADEPT-Z-a0 with core size of 8, 16, 32
- 2.47× higher *accuracy-weighted area-energy efficiency product (AAEE)* than MZI [Nat. Photon'17] and MMI [Sci. Rep.'24] arrays
- 1.03× higher AAEE than Butterfly Mesh [ASP-DAC'20]

K	Metrics	MZI [1]	Butterly [4]	MMI [10]	ADEPT-Z-a0
8	CD	0.106	0.141	0.109	0.144
	EE	9.010	9.047	9.049	9.019
	AEE	0.753	0.995	0.770	1.012
	AAEE	0.729	0.978	0.760	0.994
16	CD	0.110	0.272	0.127	0.273
	EE	16.570	18.081	17.883	18.144
	AEE	0.523	0.962	0.477	0.969
	AAEE	0.509	0.944	0.470	0.948
32	CD	0.090	0.432	0.074	0.534
	EE	26.934	34.407	32.802	36.344
	AEE	0.286	0.886	0.187	0.948
	AAEE	0.283	0.867	0.184	0.927

2-layer CNN+MNIST



ADEPT-Z Searched Design Layout Visualization (Different PTC Size K=8/16/32)

Copyright © 2022 Arizona Board of Regents

Customized PDK Adaptation

- Adapt to customized foundry PDK (with smaller PSs than GF, different PDs)
- 8.26× higher AAEE than MZI [Nat. Photon'17] and MMI [Sci. Rep.'24] arrays
- 1.04× higher AAEE than Butterfly Mesh [ASP-DAC'20]

4 variants (small-to-large) w/ customized foundry PDK

Metrics	MZI [1]	Butterfly [4]	MMI [10]	ADEPT-Z-a0	ADEPT-Z-a1	ADEPT-Z-a2	ADEPT-Z-a3
Area(O+E)	7.51+0.33	1.25+0.33	16.90+0.33	1.19+0.33	1.22+0.33	1.21+0.33	1.71+0.33
Power	223.59	219.93	219.45	218.37	218.54	218.30	218.93
Latency	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Accuracy	98.74	98.16	98.58	97.67	98.02	98.08	98.18
CD	0.653	3.242	0.297	3.370	3.301	3.322	2.507
EE	22.899	23.280	23.330	23.446	23.429	23.454	23.387
AEE	2.919	14.744	1.354	15.434	15.105	15.215	11.450
AAEE	2.882	14.473	1.336	15.074	14.806	14.923	11.242

Copyright © 2022 Arizona Board of Regents

Generalize to Different Tasks/Models

- Search PTC on 2-layer CNN + MNIST → Train ONN on different tasks/models
- Average of 1.6× higher AAEE than manual PTC designs

Model	Dataset	MZI [1]	Butterfly [4]	MMI [10]	ADEPT-Z-a0
CNN	FMNIST	0.472	0.852	0.432	0.853
VGG8	CIFAR10	0.429	0.744	0.382	0.769
ResNet20	SVHN	0.491	0.884	0.445	0.890

AAEE↑

- **Future direction in CAD for optical computing**
  - Search beyond this layer-wise coherent PTC design template
  - Architecture, interconnect, and PTC topology co-search

Copyright © 2022 Arizona Board of Regents

Thank you!
Q & A?

ADEPT-Z: Zero-Shot Automated Circuit Topology Search for Pareto-Optimal Photonic Tensor Cores

Ziyang Jiang[1], Pingchuan Ma[1], Meng Zhang[2], Rena Huang[2], Jiaqi Gu[1*]
[1]Arizona State University, [2]Rensselaer Polytechnic Institute

Contact Jiaqi Gu: jiaqigu@asu.edu

Open-Source TorchONN Toolchain

Opensource Code

Automating optical AI hardware design toward productivity

Arizona State University

ASU Center for Semiconductor Microelectronics (ACME)

SCOPE

Copyright © 2022 Arizona Board of Regents