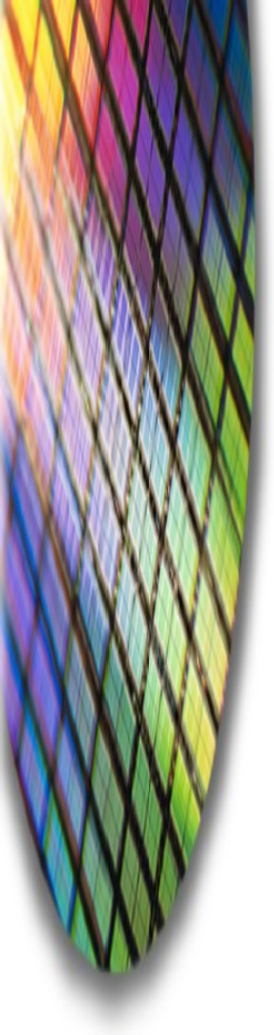


A3D3 Seminar, 11/7/2022

Closing the Virtuous Cycle of AI for IC and IC for AI

David Z. Pan

ECE Department, UT Austin

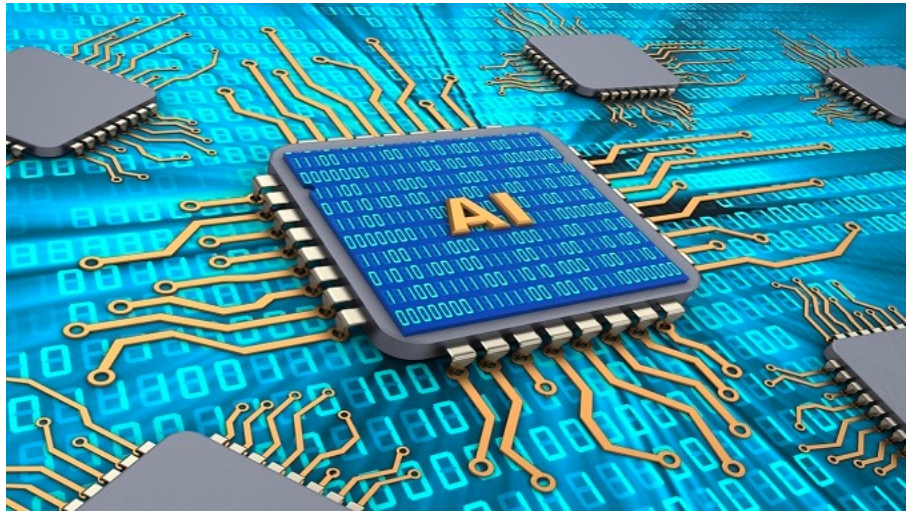
https://www.ece.utexas.edu/~dpan

AI and the "ABC" Behind

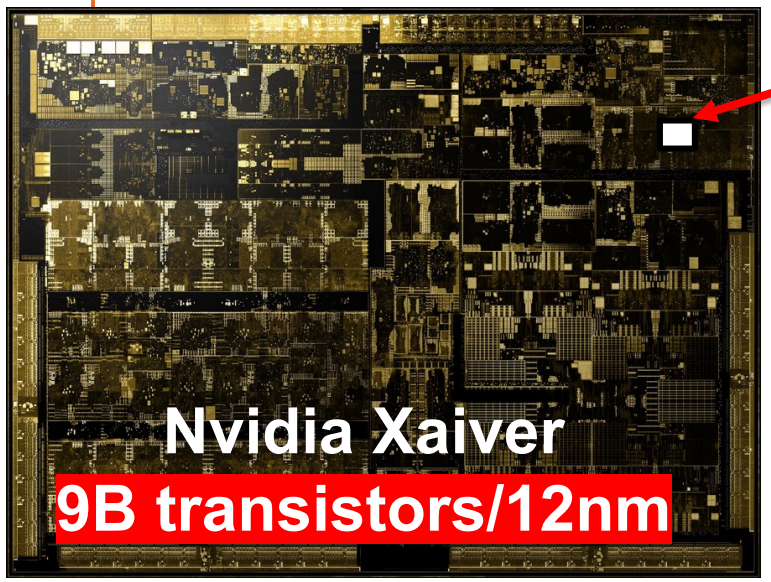
- Algorithms: CNN, RNN, GAN, …

- Big data

- Chips: CPU, GPU, ASIC, FPGA, and dedicated AI accelerators

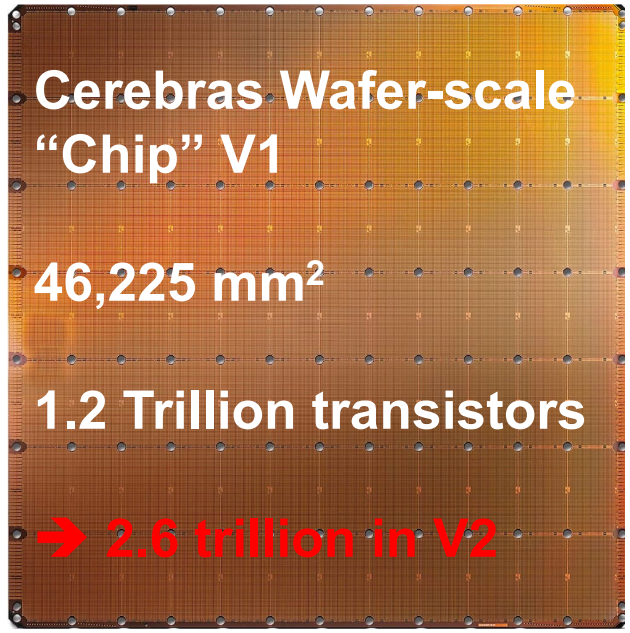


Nanometer IC Design/Manufacturing Complexity



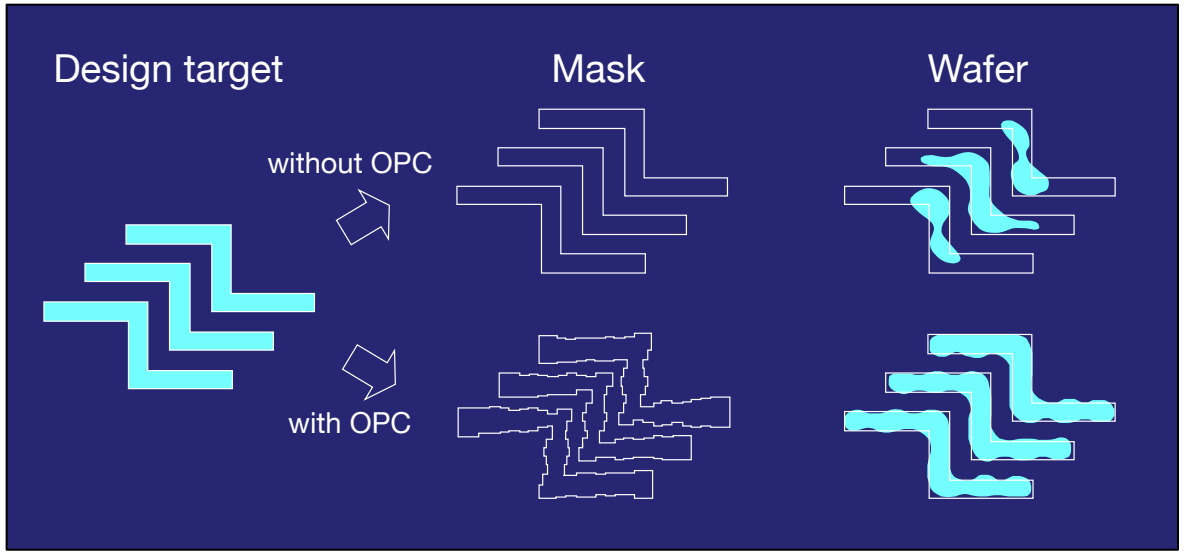
Divide large chip into smaller partitions, e.g., 1~2M cells each

Still, 1 backend iteration for one partition could take days!

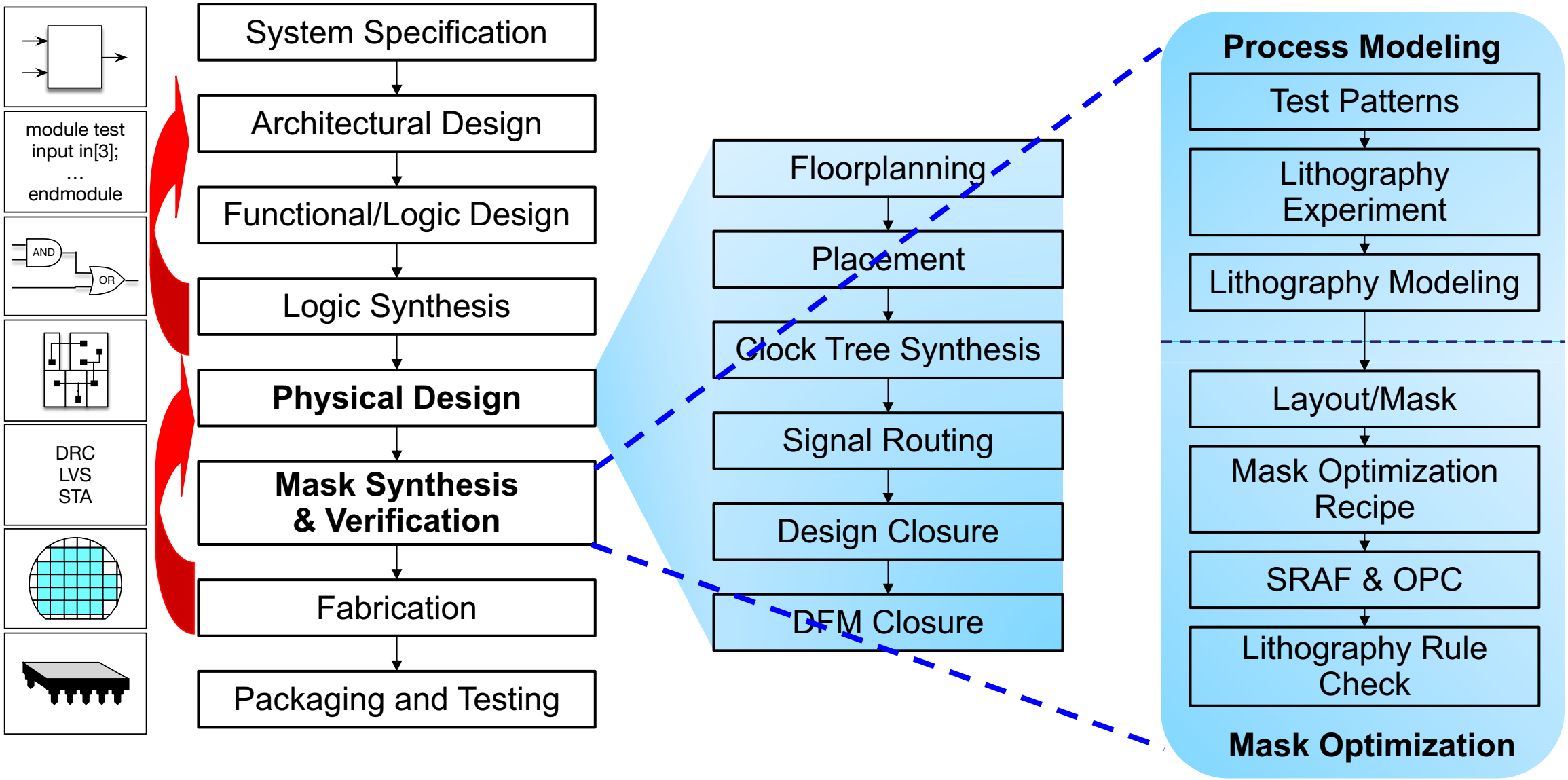


8,000 Engineer-Year!

What you see (at design) is not (necessarily) what you get (at fab)!



IC Design/Manufacturing Flow



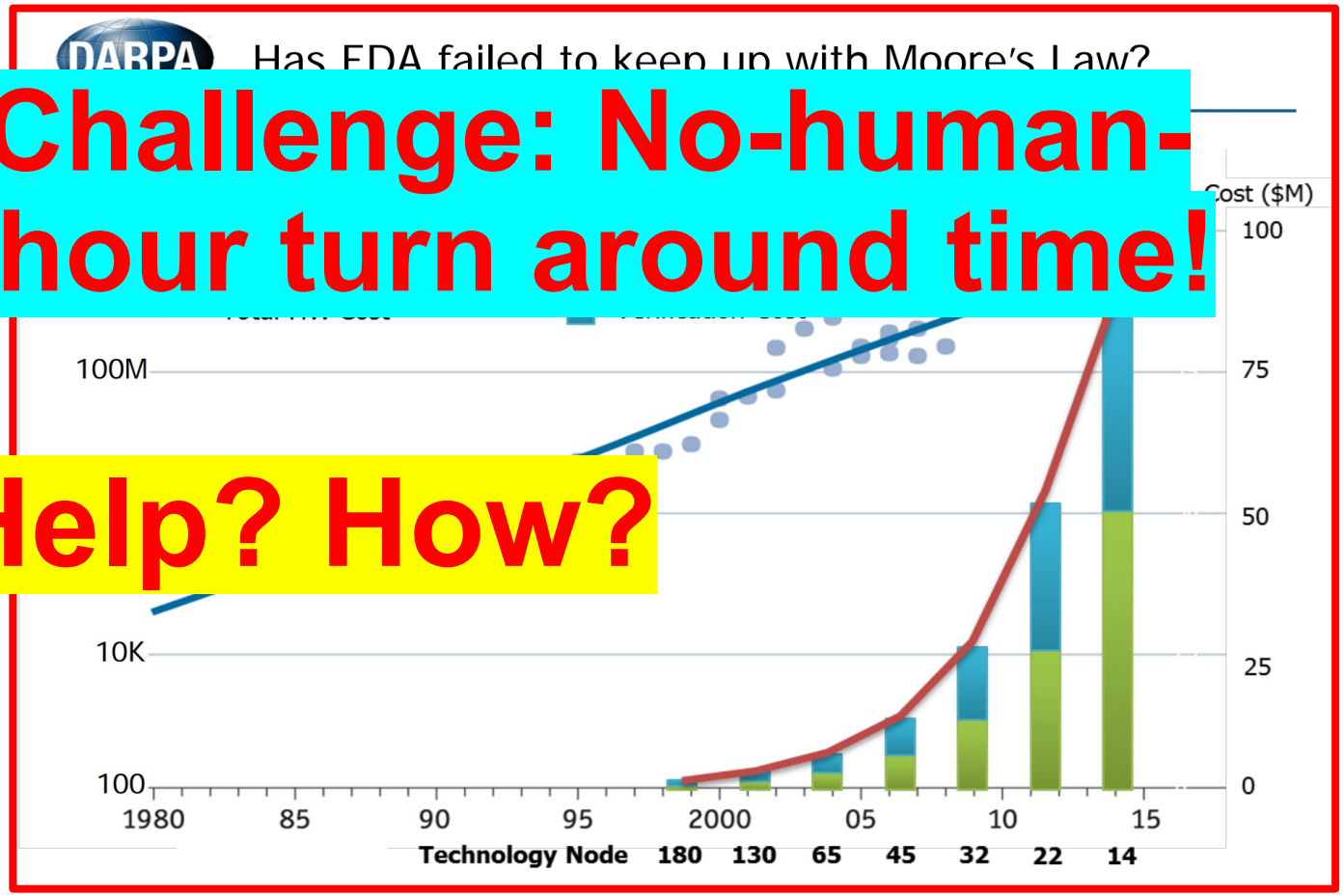
Nanometer Design/Manufacturing Challenges

- Performance/Power/Area (PPA)
- Manufacturability/Yield
- Reli
- Security
- Design cost
- …
- DARPA ERI ($1.5B) IDEA/POSH “Silicon Compiler 2.0” ($100M)

A. Olofsson, DARPA, ISPD-2018 Keynote

DARPA Grand Challenge: No-human-in-the-loop, 24-hour turn around time!

Can AI Help? How?

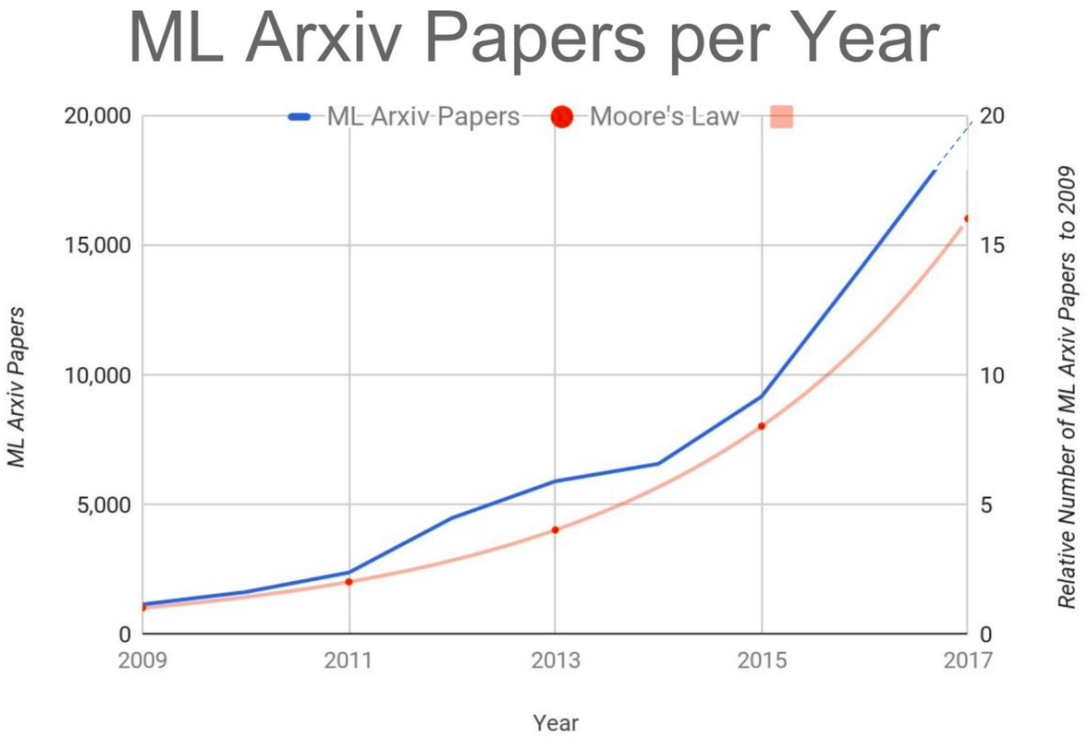


AI – IC Interactions

Two key themes

- **AI for IC**
  - How to leverage AI techniques to enable agile and intelligent IC design
  - Equivalent scaling of Moore's law
  - Democratizing IC and EDA R&D
- **IC for AI**
  - Customized IC/FPGA for AI applications
  - Efficient/hardware aware ML

Interestingly …



Closing the virtuous cycle!

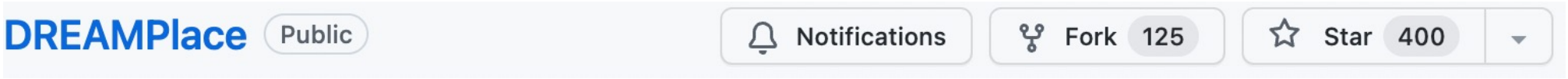
Outline

- Introduction
- **AI for IC**
- IC for AI
- Conclusion

Case Study 1

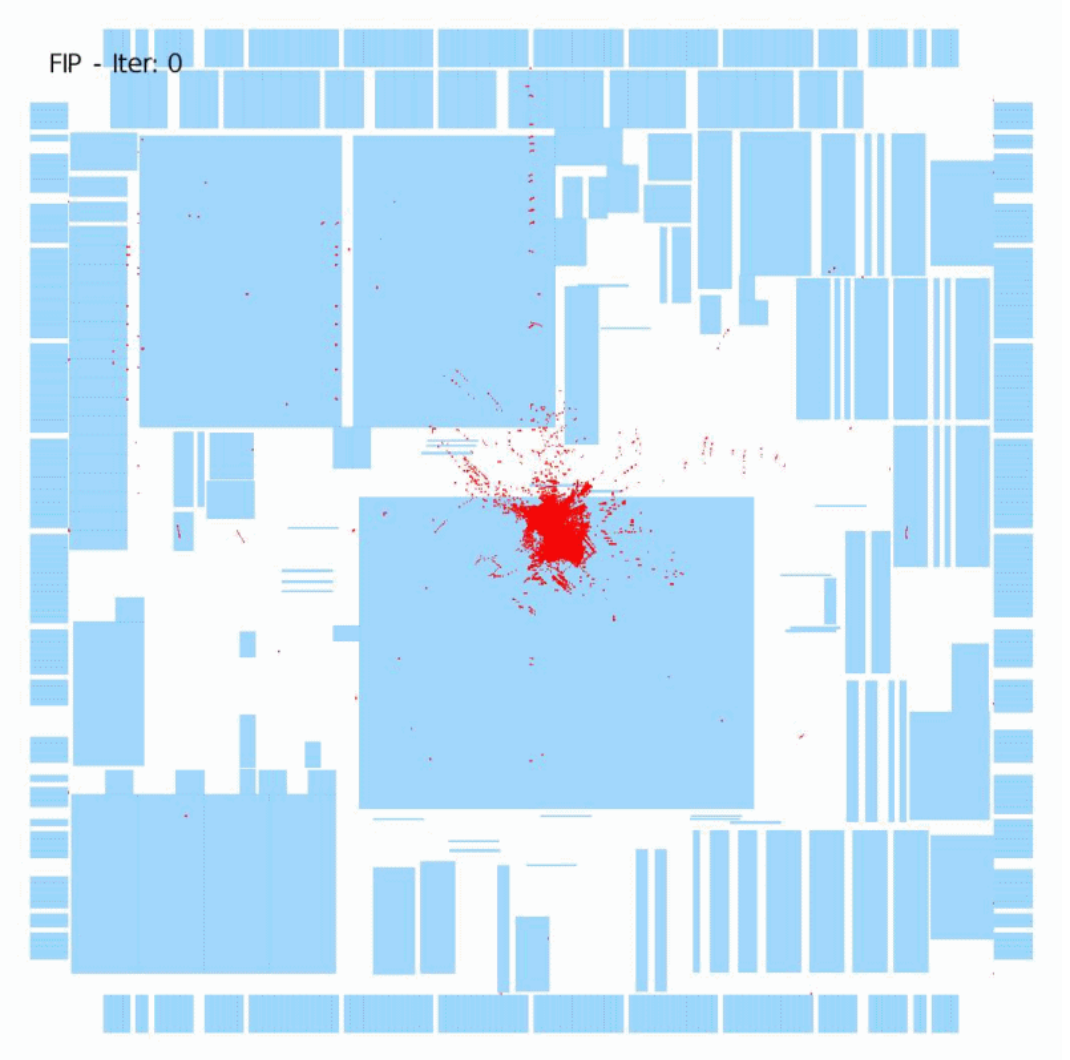
DREAMPlace: Deep Learning Toolkit-Enabled GPU Acceleration for Modern VLSI Placement [Lin+, DAC'19 Best Paper Award; IEEE TCAD 2021 Donald O. Pederson Best Paper Award]

Source code release: https://github.com/limbo018/DREAMPlace
Widely used by industry (Google, Nvidia, Intel, …) and academia



Challenges of VLSI Placement

- A classical NP-hard problem!
- Have to deal with huge designs: 10M+ cells in modern ICs
- Plays a central role in IC design closure as it is in the middle of the entire design flow
  - Placement determines the interconnect to the first order
  - Modern designs are interconnect-centric



Courtesy RePlAce from UCSD

Typical SOTA Nonlinear Placement Algorithm

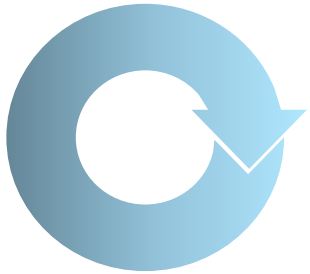
$$\min_{\mathbf{x},\mathbf{y}} \quad \sum_{e \in E} \mathrm{WL}(e;\mathbf{x},\mathbf{y}),$$

$$s.t. \quad D(\mathbf{x},\mathbf{y}) \leq t_d$$

Objective of nonlinear placement

$$\min \quad (\sum_{e \in E} \mathrm{WL}(e;\mathbf{x},\mathbf{y})) + \lambda D(\mathbf{x},\mathbf{y})$$

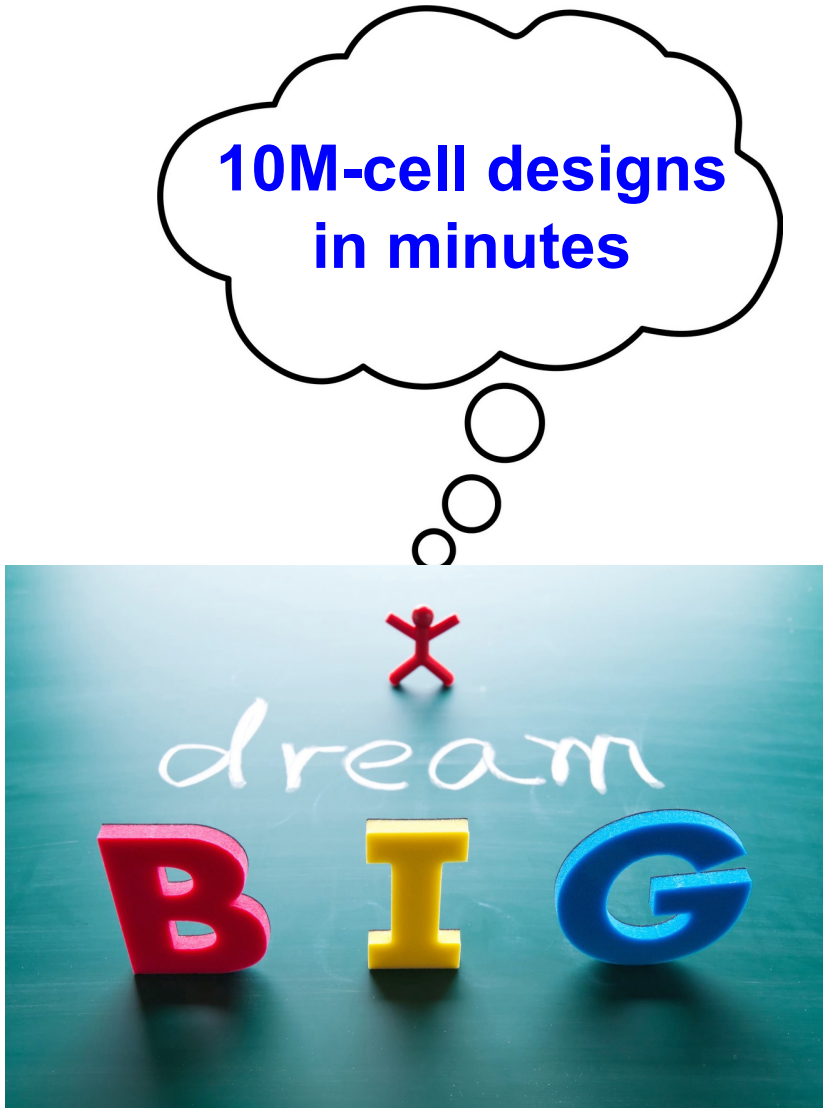
Wirelength Density



Huge development effort and runtime for high-quality placement of modern ASIC/SoC designs

What is your Dream Placement Engine?

- ✓ Best quality: wirelength ➔ congestion, timing, power, …
- ✓ Ultrafast: placement is at the center of entire design flow ➔ faster design turn-around-time
- ✓ Low development overhead: ➔ from 1 year to a month?
- ✓ Extensible: to new algorithms and acceleration techniques

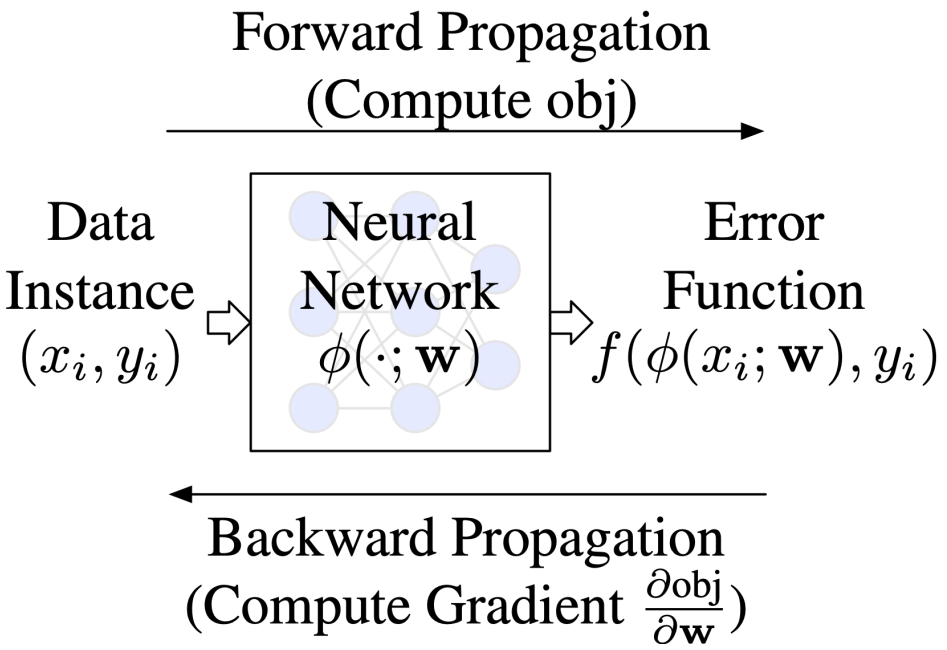


DREAMPlace Strategies

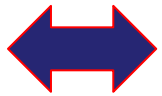
- We propose a **novel analogy** by casting the nonlinear placement optimization into a neural network training problem
- Greatly leverage deep learning **hardware** (GPU) and open-source **software** toolkits (e.g., PyTorch)
- Enable ultra-high parallelism and acceleration while getting state-of-the-art results

Analogy Between NN Training and Placement

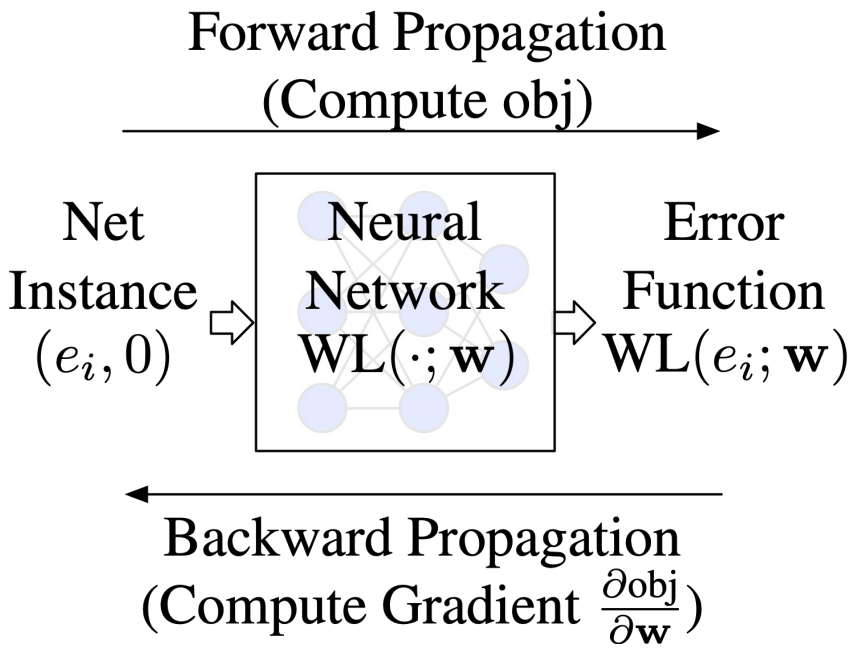
$$\min_{\mathbf{w}} \sum_{i}^{n} f(\phi(x_i; \mathbf{w}), y_i) + \lambda R(\mathbf{w})$$



Train a neural network



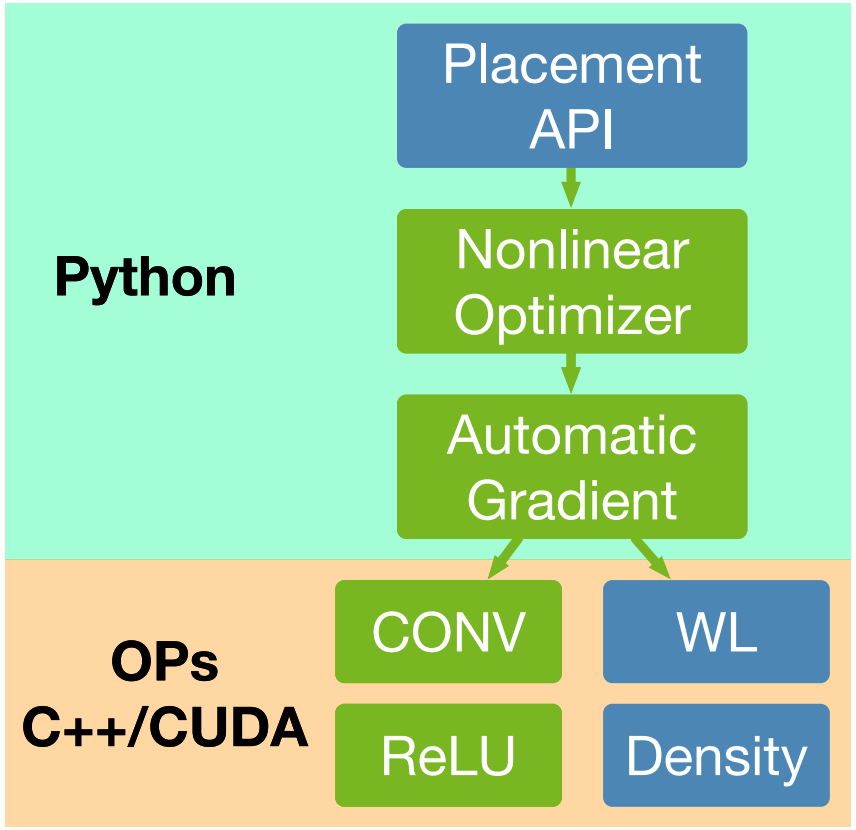
$$\min_{\mathbf{w}} \sum_{i}^{n} \text{WL}(e_i; \mathbf{w}) + \lambda D(\mathbf{w})$$



Solve a placement

DREAMPlace Architecture

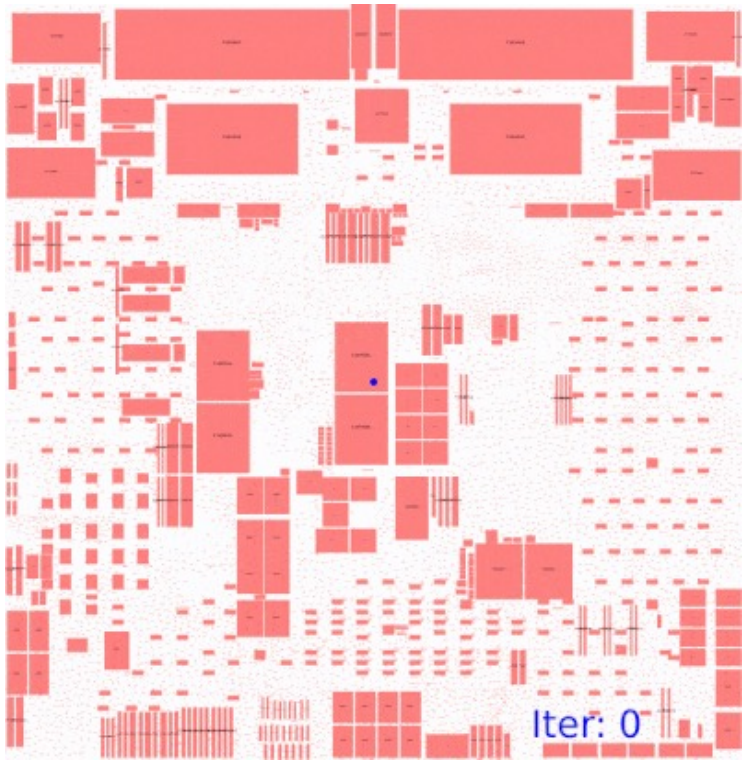
Leverage highly optimized deep learning toolkit



DREAMPlace architecture

Match RePlAce [Cheng+, TCAD18]

Nesterov's Method



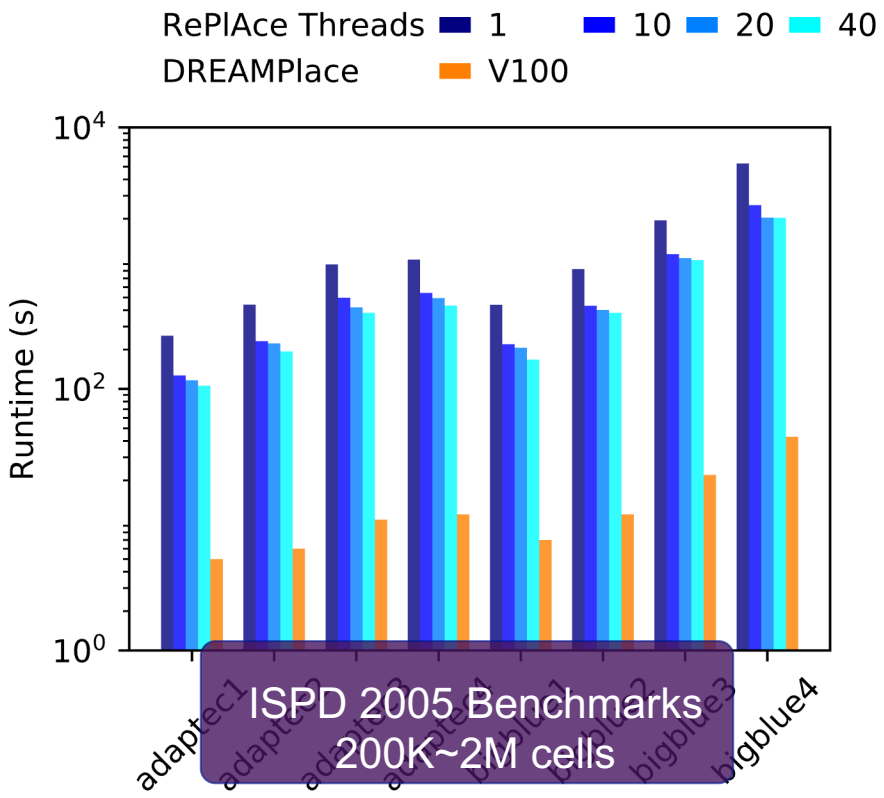
Global Placement Result Comparison

RePlAce [Cheng+, TCAD'18]
- CPU: 24-core 3GHz Intel Xeon
- 64GB memory allocated
- Current state-of-the-art

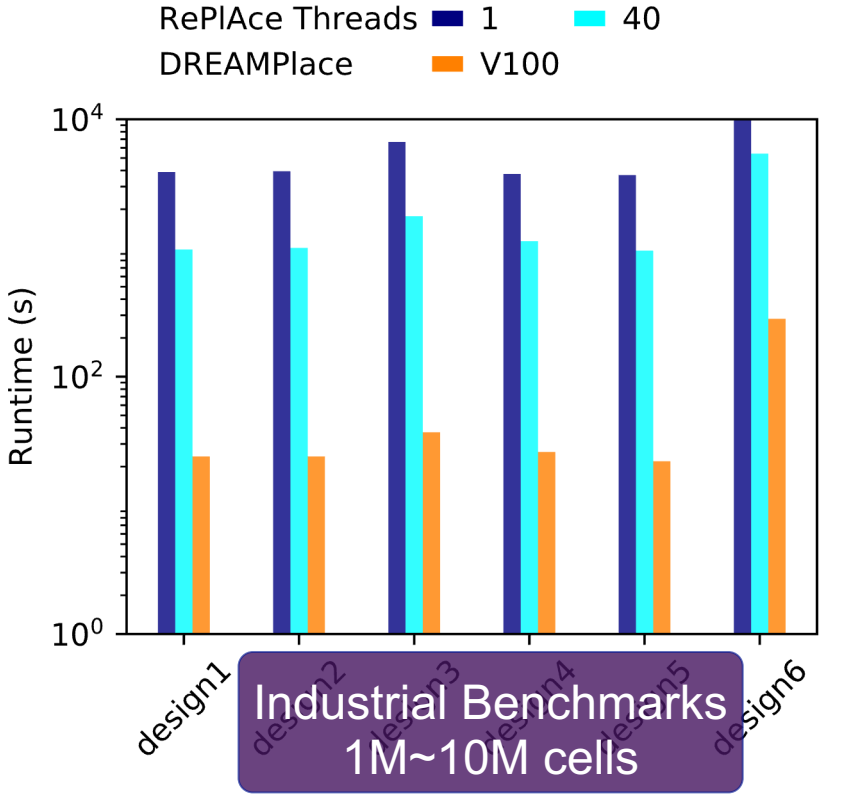
DREAMPlace [Lin+, DAC'19]
- CPU: Intel E5-2698 v4 @2.20GHz
- GPU: 1 NVIDIA Tesla V100
- Single CPU thread was used

Same placement quality of results!

34× speedup by DREAMPlace



43× speedup by DREAMPlace



10M-cell design finishes in min, instead of 3+ hrs

Dreams for DREAMPlace

✓ Best Quality

Match state-of-the-art quality

✓ Ultrafast

Over 30x speedup
10M-cell design
3h➔5min

DREAM
Comes True

Easy algorithm innovation
Acceleration innovation

✓ Extensible

Coding effort 1yr ➔ 2 mon
Leverage existing toolkits

✓ Low Development Overhead

Beyond DREAMPlace

New Solvers

SGD, ADAM, etc.

[TCAD'21]

New Objectives/constraints

Routability, timing, fence…

DREAMPlace 2.0, 3.0, …

DREAM BIGGER

Macro placement w/ Google [MLCAD'21];
DREAMPlaceFPGA [ASPDAC'22]; Gate sizing…

Other CAD Problems

Multi-GPU,
Distributed computing,
Mixed precision,
…

New Accelerations

Case Study 2

MAGICAL: Machine Generated Analog IC Layout

As part of DARPA ERI (IDEA/POSH) effort

Open source MAGICAL (v1.0) released

https://github.com/magical-eda/MAGICAL

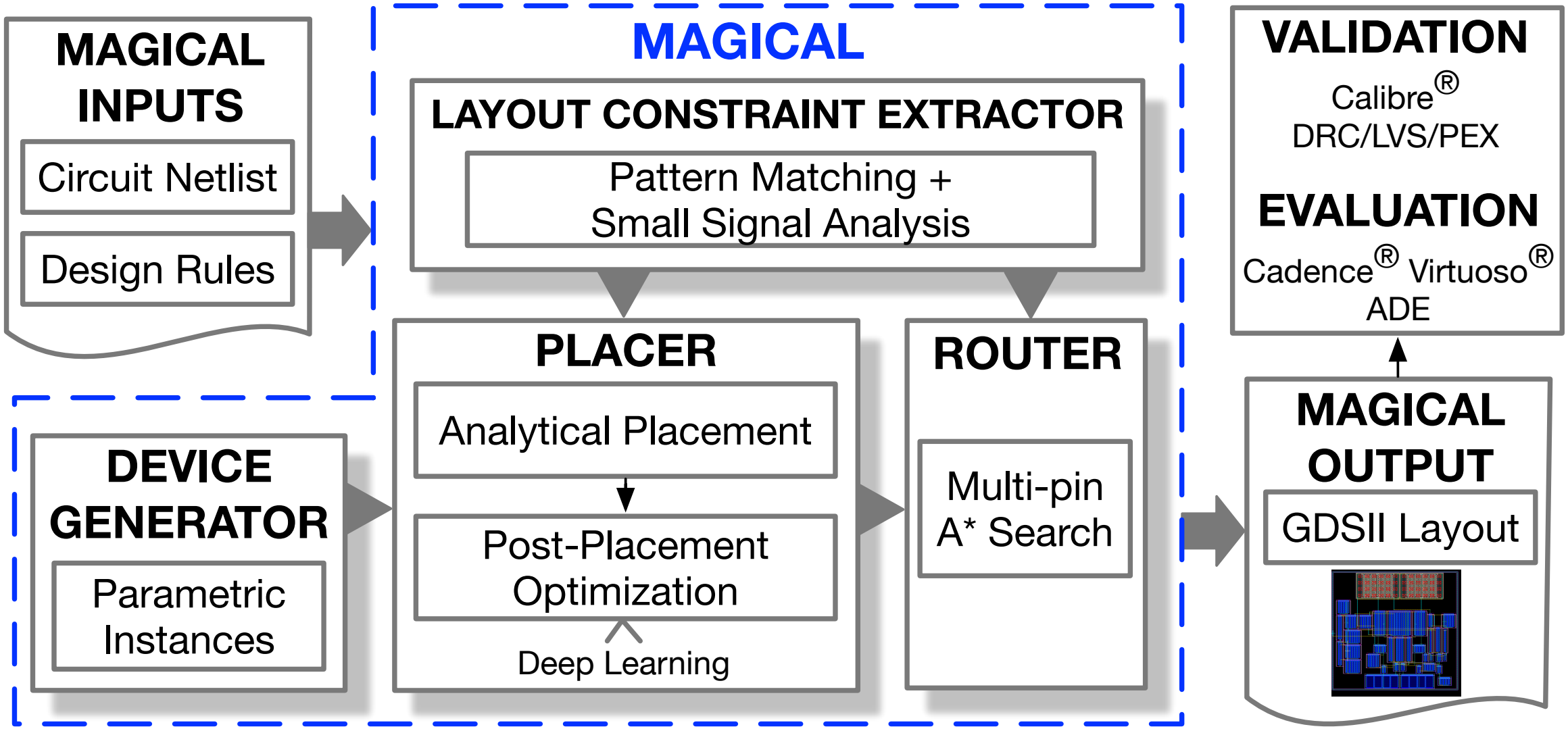
Analog IC Layout

- DREAMPlace mainly for digital IC
- Analog IC to interface with outside world
- Analog IC layout design still mostly **manual**
  - Very tedious and error-prone
  - Prior DA not as successful as that in digital IC

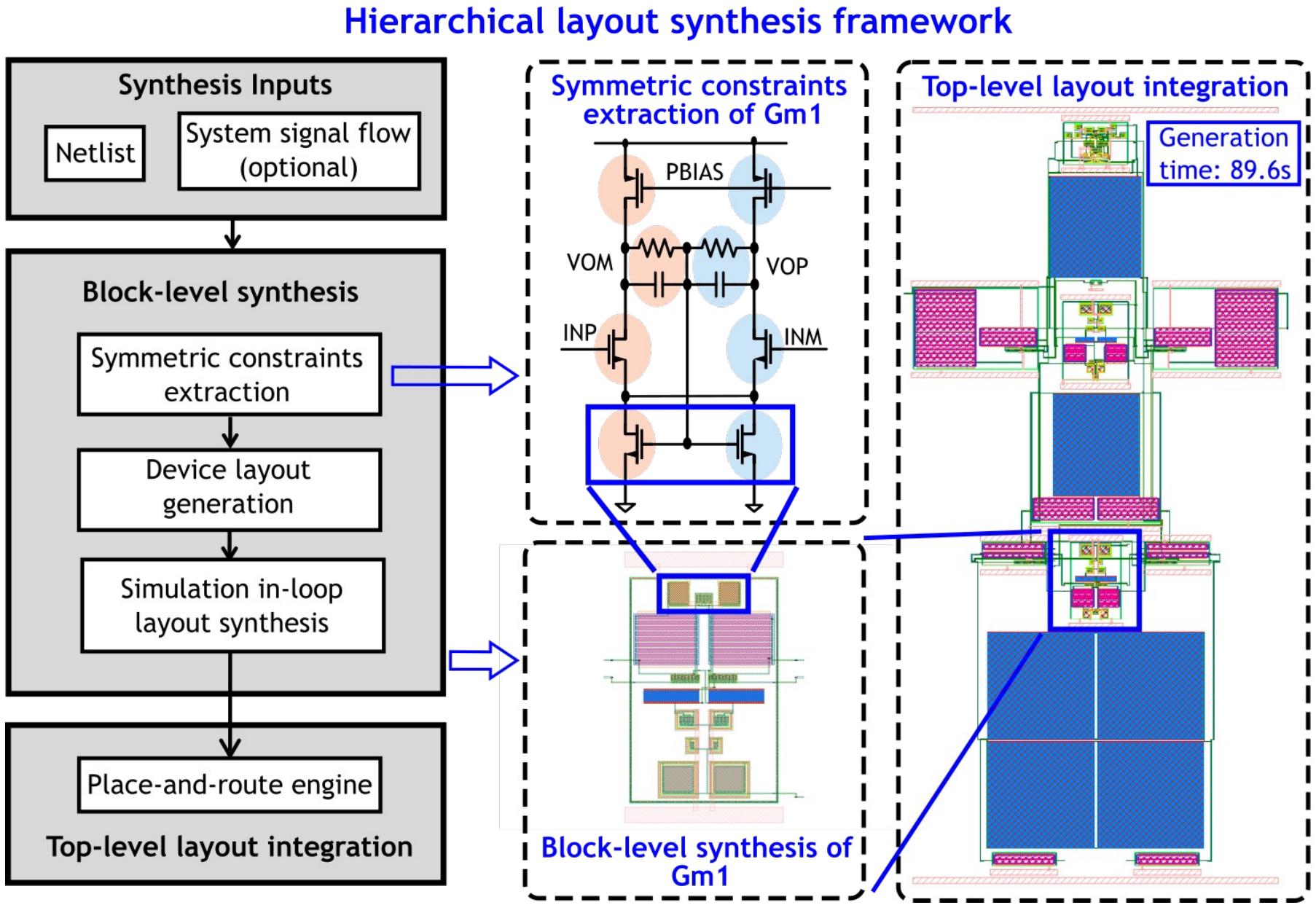
MAGICAL Mission:

- Develop a fully-automated analog layout system, leveraging human and machine intelligence
- Promising results [ISPD'19, DAC'19, ICCAD'19, ASPDAC'20, DATE'20, DAC'20, ICCAD'20, D&T'20, JoS'20, CICC'21, DAC'21, ICCAD'21, ASPDAC'22, DATE'22, ISPD'22, ICCAD'22]

MAGICAL Layout System Framework



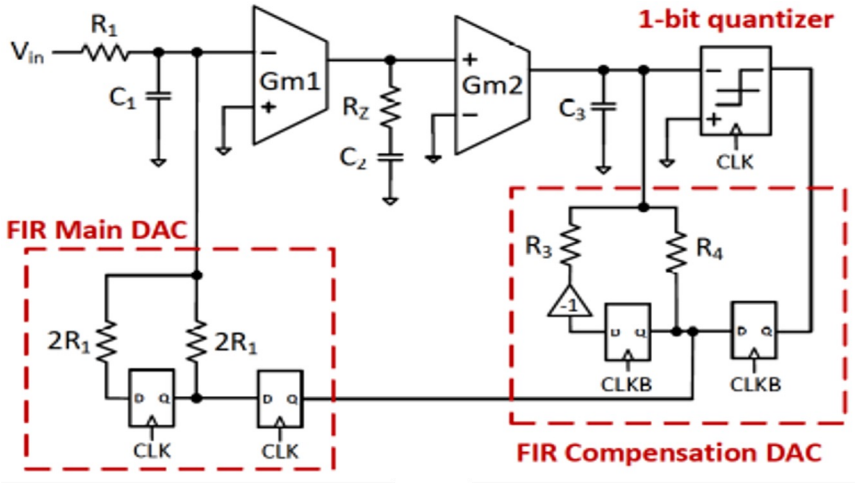
MAGICAL 1.0 Hierarchical Framework [Chen+, CICC'21]

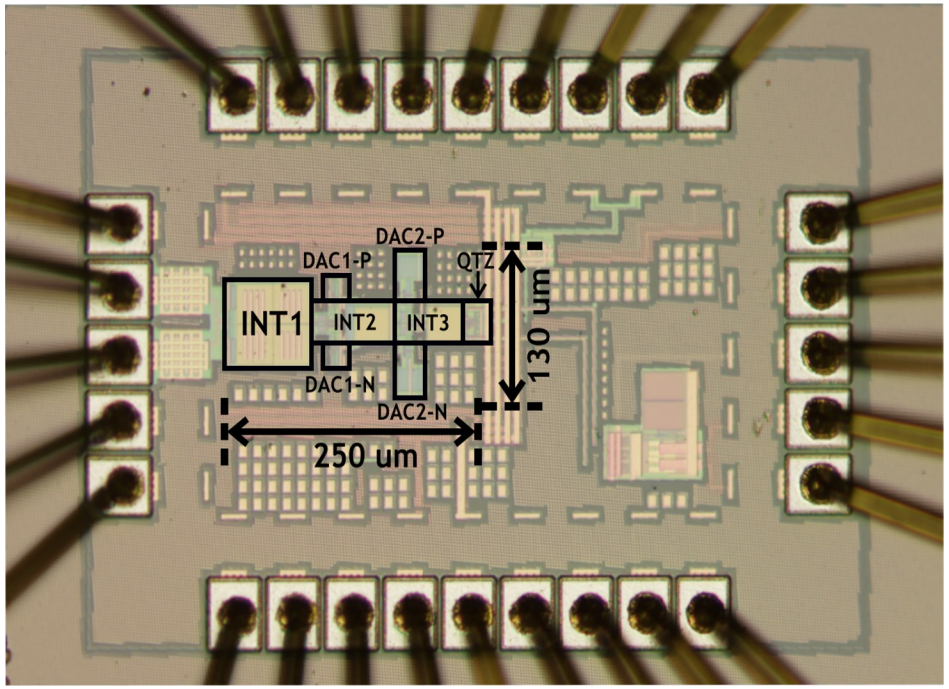


MAGICAL 1.0 Tapeout

[Chen+, CICC'21]

- 1GS/s 3rd-order high-performance continuous time ΔΣ modulator
- Include various sub-block types
  - Three integrators: one passive, two active
  - Two FIR-based feedback DACs
  - One comparator
  - + Digital logics
- TSMC 40nm
- SOTA performance cf. the original manual design [IEEE SSC-L'20]





Comparison with SOTA CTDSM ADCs

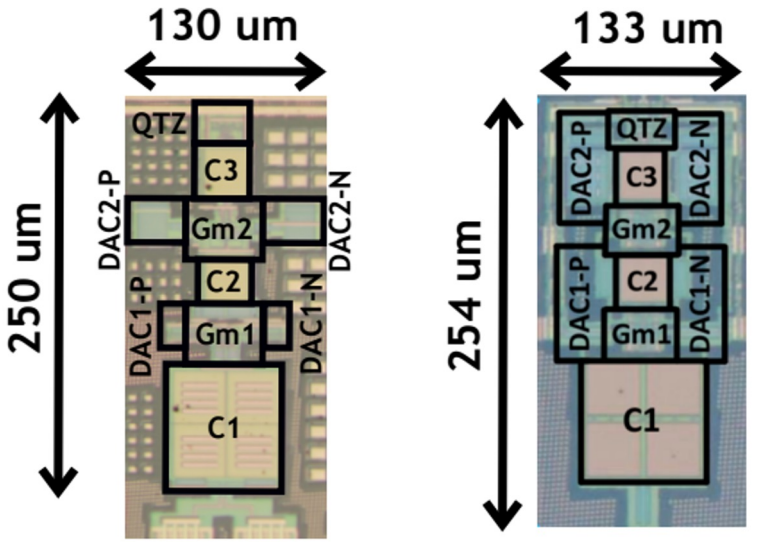
- MAGICAL 1.0 layout even slightly outperforms manual layout (SSCL'20) in power, performance, and area

	JSSC-16 Weng	SSCL-20 Mukherjee[1]	CICC-19 Li	This work[1]
Architecture	CTΔΣM	CTΔΣM	VCO-CTΔΣM	**CTΔΣM**
Layout synthesized	✗	✗	✓	✓
Universal synthesis framework	N/A	N/A	✗	✓
Hierarchical flow	N/A	N/A	✗	✓
Constraint generation	N/A	N/A	✗	✓
Order	4th	3rd	1st	**3rd**
Process [nm]	28	40	40	**40**
Area [mm²]	0.1	0.034	0.01	**0.033**
Fs [MHz]	320	1024	600	**1024**
Supply [V]	1.1/1.2	1.2	1.1	**1.2**
Power [mW]	4.2	0.79	1.08	**0.77**
BW [MHz]	10	5	4	**5**
SFDR [dB]	94.2	82.6	75	**80.8**
SNDR [dB]	74.4	65.6	68.8	**67.4**
FoMw[2] [fJ/conv-step]	49.3	51	60	**40.2**
FoMs[3] [dB]	174.5	163.6	164.3	**165.5**

[1]Same schematic

[2]$FoM_W = Powe$

[3]$FoM_S = SNDR$



MAGICAL 1.0

O(Month)

O(Min)

[Chen+, CICC'21]

MAGICAL Extension: OpenSAR

[Liu+, ICCAD'21]

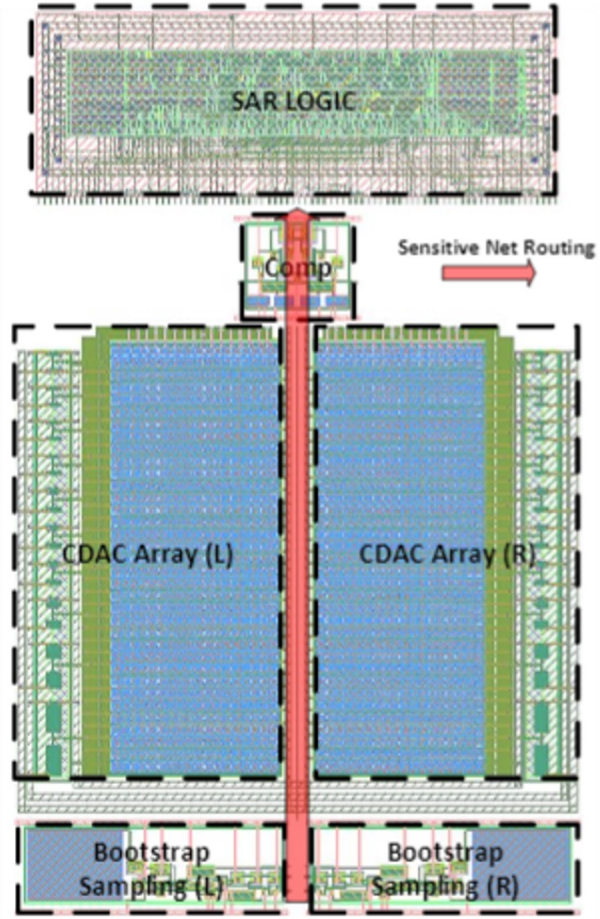
- End-to-end SAR ADC compilation

Digital APR

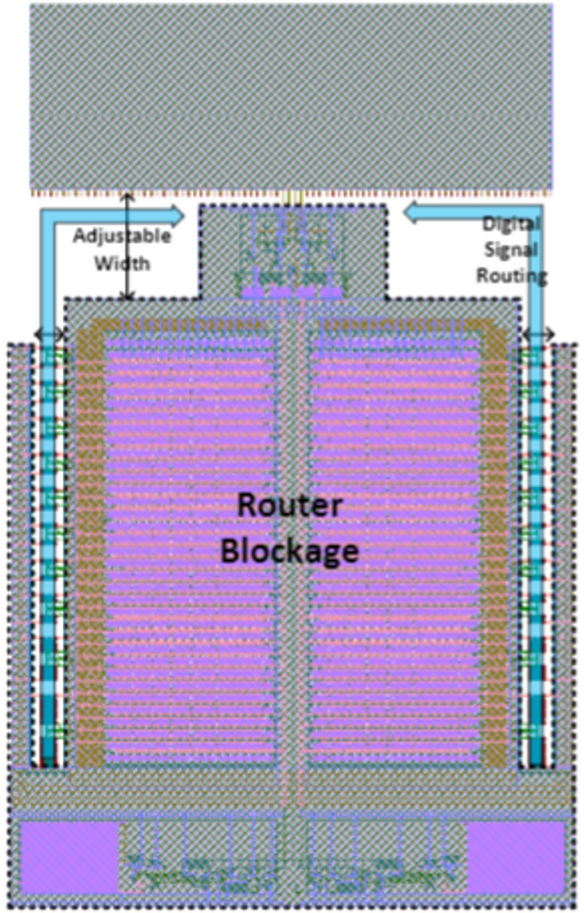
MAGICAL

Template-based Generation

MAGICAL

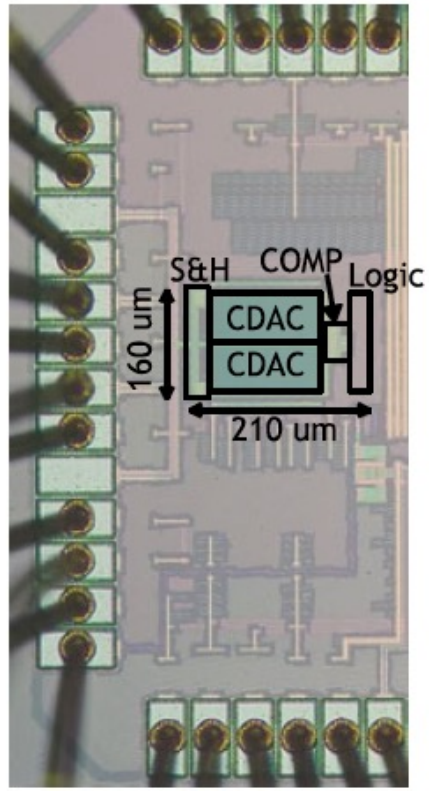


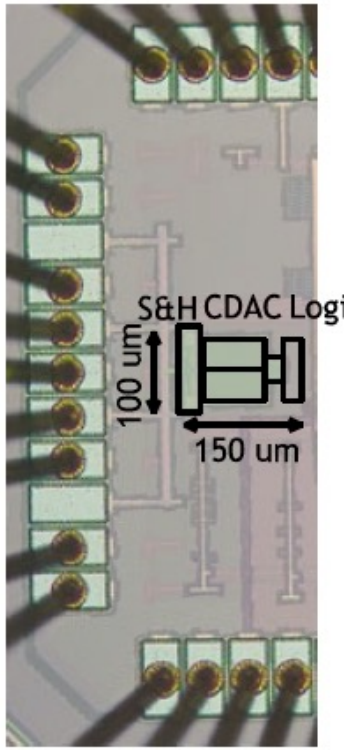
Floorplan



Route planning

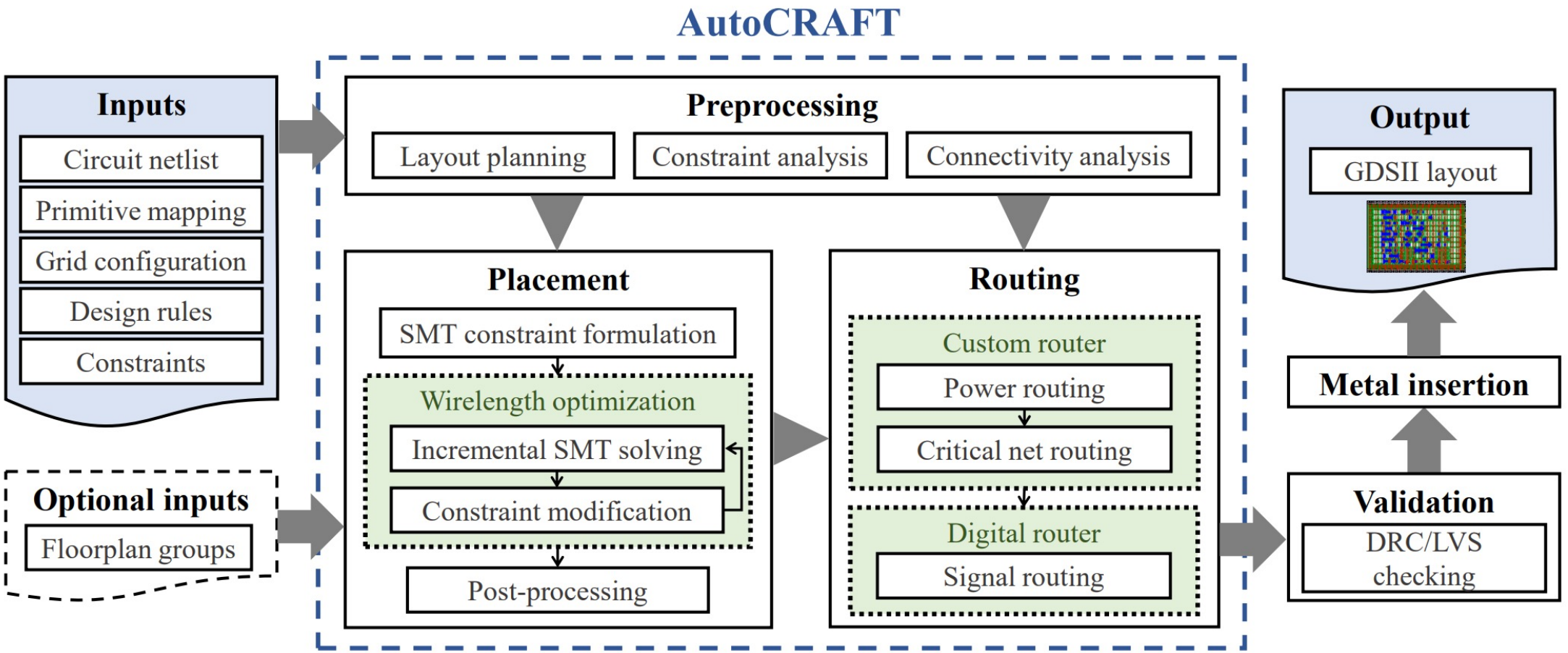
Tape-out validated
TSMC 40nm





MAGICAL Extension: AutoCRAFT [Chen+, ISPD'22]

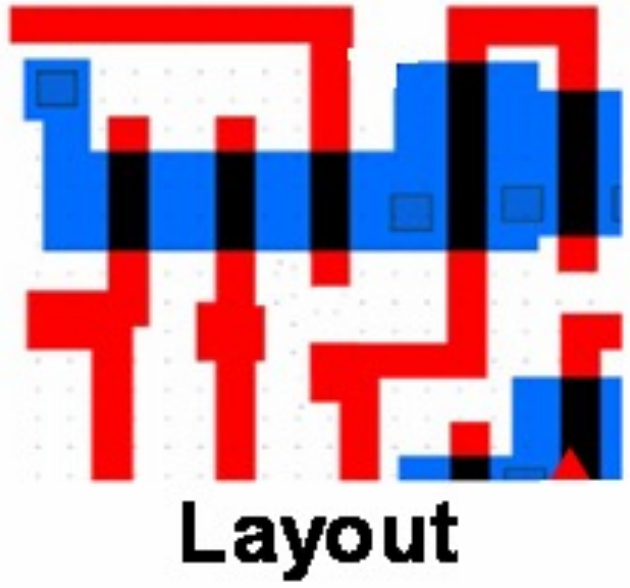
- Tech-agnostic **FinFET** layout style using primitives (w/ Nvidia)
- Auto custom layout generation ➔ Very promising results obtained

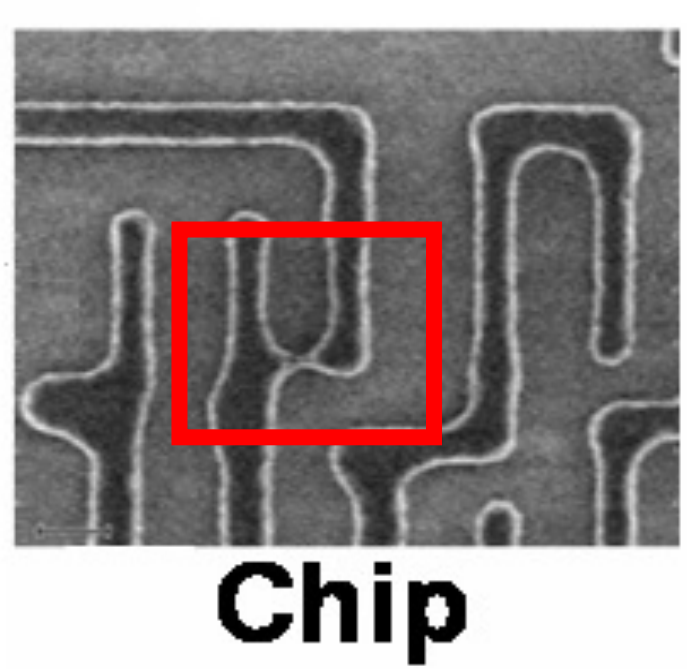


Case Study 3

AI for IC Manufacturability, Reliability, Security

Bottleneck in IC Manufacturing: Lithography



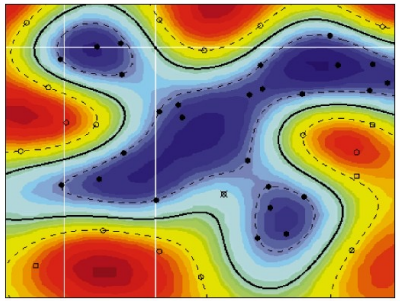


- What you see (at design) is **NOT** what you get (at fab)
- Need to make sure design is manufacturable with high yield
- Litho-simulations are **extremely** CPU intensive

Lithography Hotspot Detection

Question 1: Without going through detailed litho-simulations, can we directly predict lithography hotspot to avoid poor yield?

- Our work [Ding+, ICICDT 2009 Best Paper] is **among the first** to use machine learning (SVM) for litho-hotspot detection
  - Very active research topic in the last 12+ years
  - **Inspired ICCAD 2012 CAD Contest**, run by Mentor Graphics
  - Meta-classification combining ML and PM [Ding+, ASPDAC'12 BPA]
  - Deep neural network [Yang+, DAC'17]
  - Big data vs. small data: transfer learning, active learning, semi-supervised learning [Lin+, ISPD'18], [Chen+, ASPDAC'19] …
  - Litho-GPA: confidence estimation [Ye+, DATE 2019]
  - ……

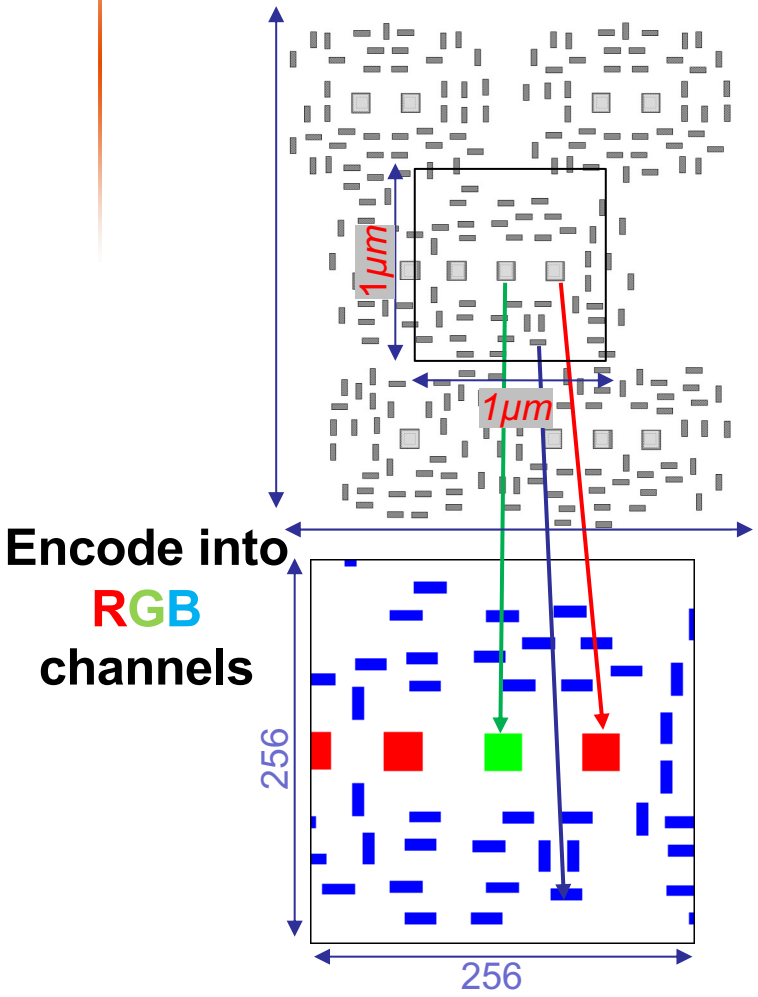


LithoGAN: End-to-End Lithography Modeling with Generative Adversarial Networks [Ye+, DAC'19 Best Paper Finalist]

Question 2 (much harder): Without going through litho-simulations, can we directly get printed images?

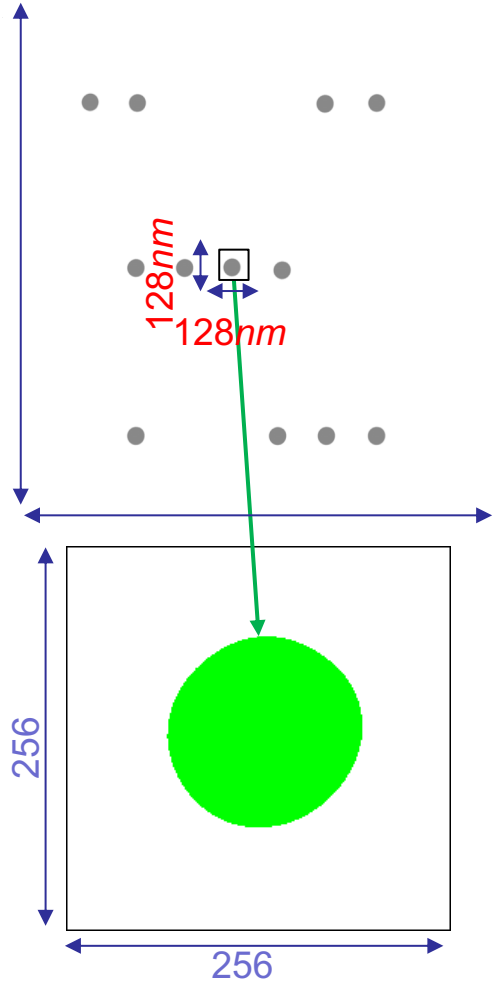
Image Translation for Litho Modeling

[Ye+, DAC'19]



Expensive Litho Simulation

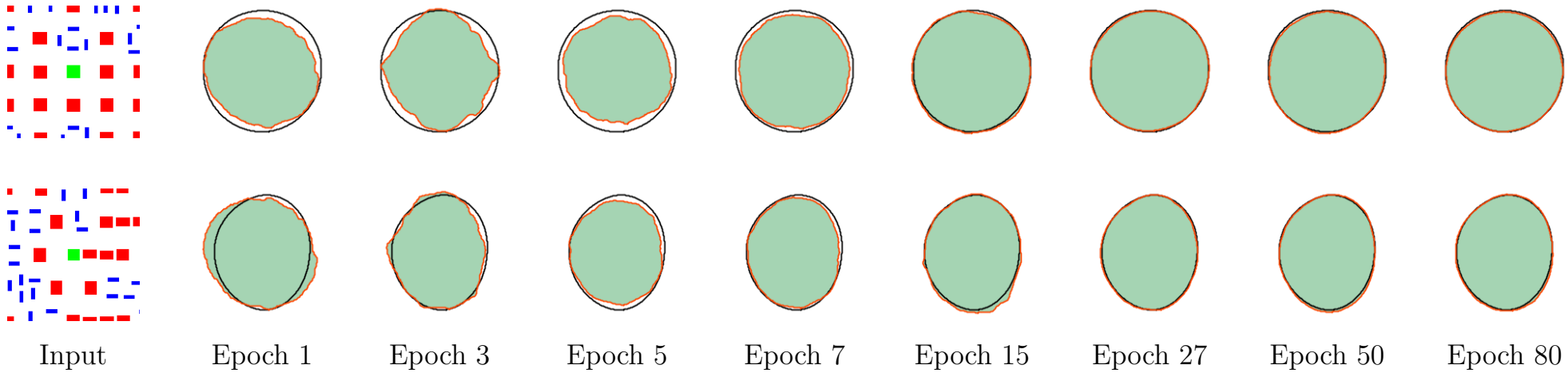
This is now a (modified) **image translation** task



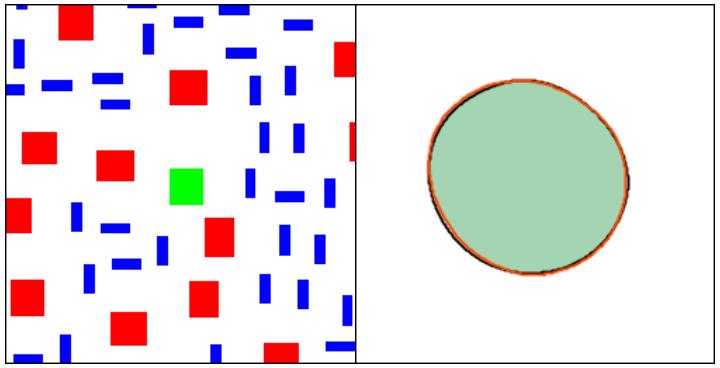
- Different elements encoded on different image channels
- Resist pattern zoomed in for high-resolution/accuracy

LithoGAN Results

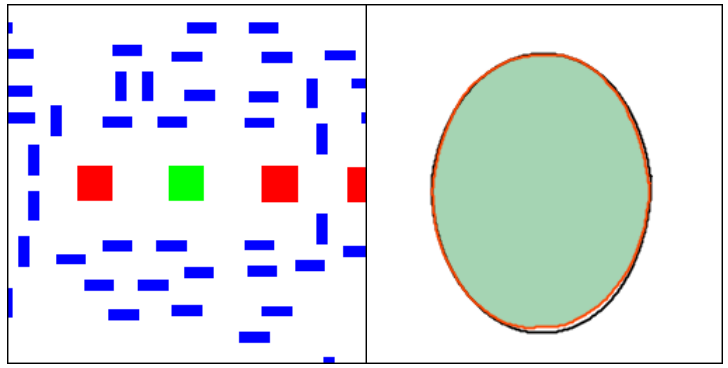
[Ye+, DAC'19]



Model advancement progress



Input LithoGAN output



Input LithoGAN output

LithoGAN is **1800x** faster than rigorous simulations, with acceptable error (in consultation with industry)

LAPD

Another LAPD

- To **bridge** design and manufacturing ➔ **L**ithography **A**ware **P**hysical **D**esign (**LAPD**)
  - Litho Hotspot Detection
  - Litho Hotspot Correction
- My group has made many seminal contributions in **LAPD** ⬆️
- **LithoGAN** opens new directions with tremendous potential
- **Similar principles apply to other EDA (reliability, 3D-IC, …)**

Detection

Correction

Bridge Design/Manufacturing for Security

- IC supply chains of design, manufacture, test, package, ...

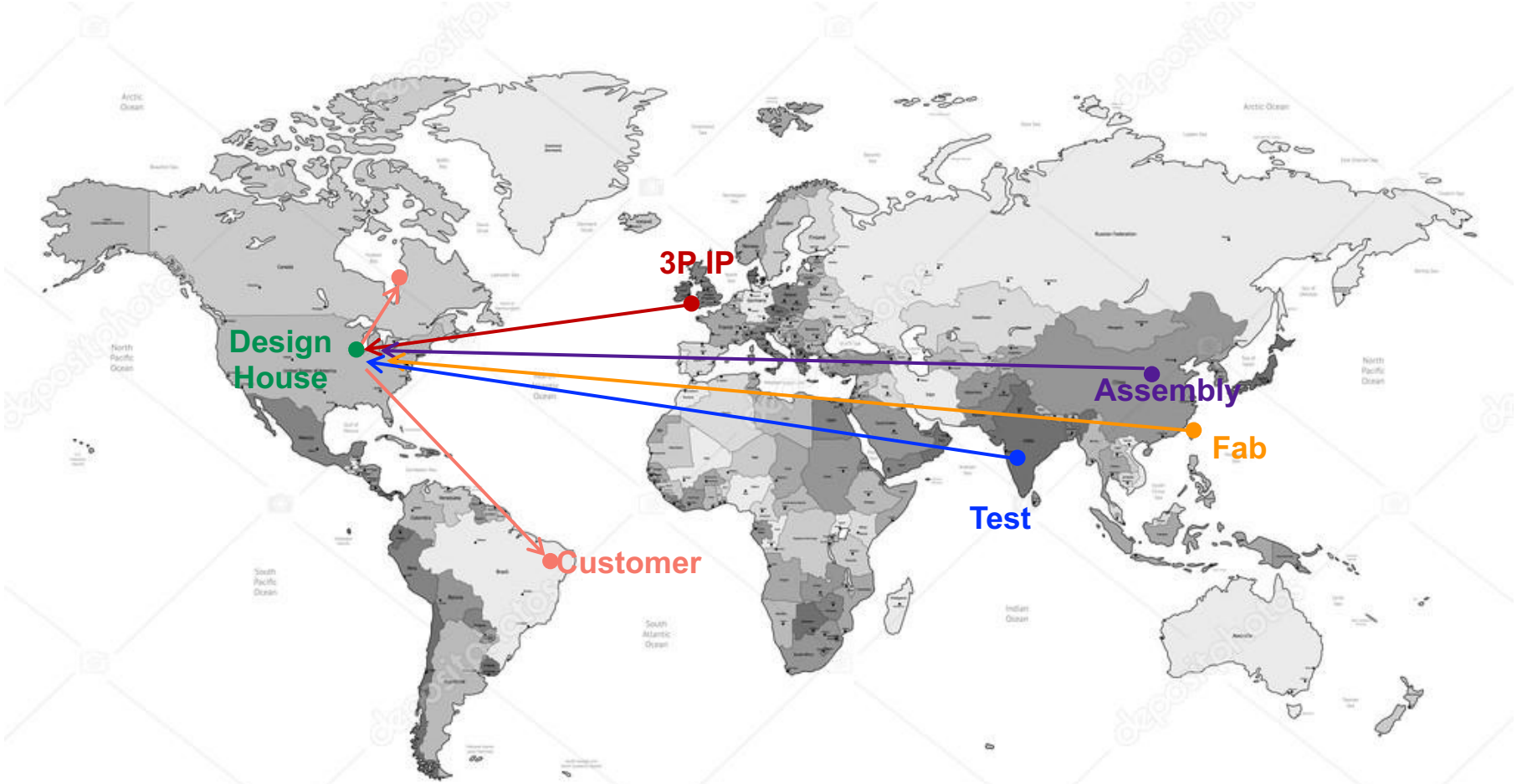
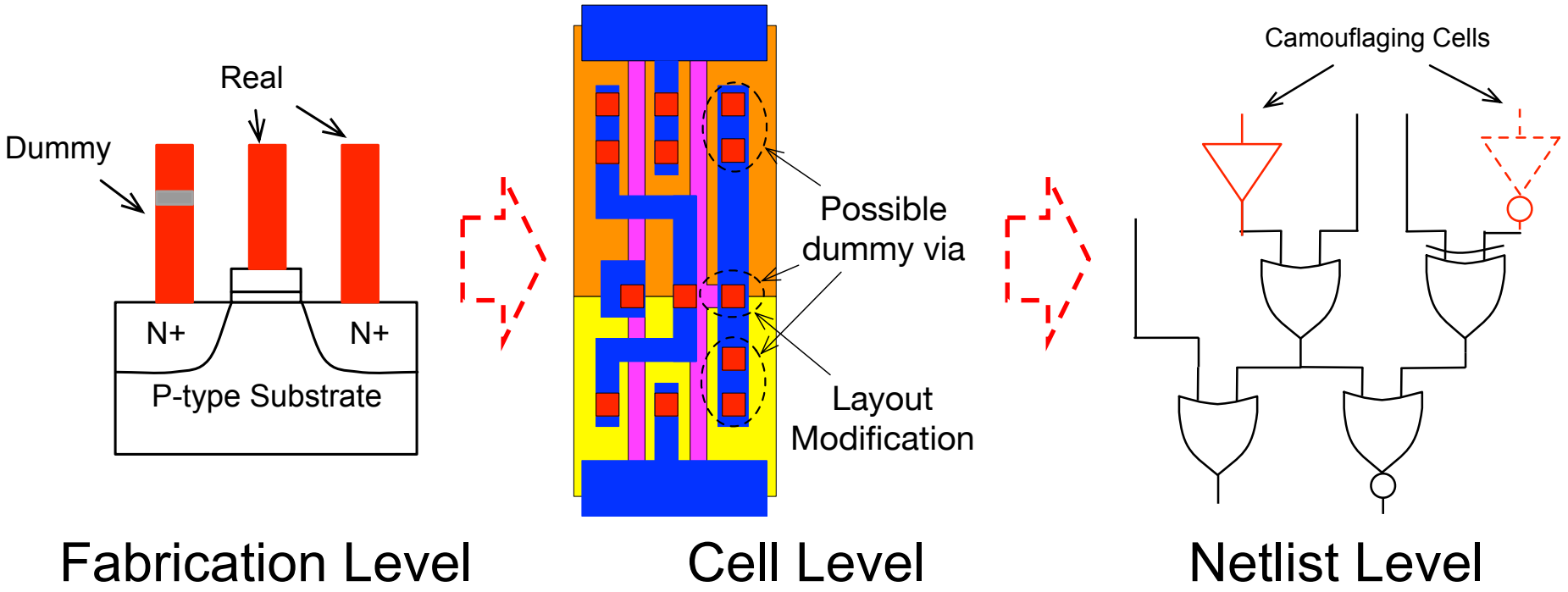


Image source: https://depositphotos.com/2801291/stock-illustration-gray-detailed-world-map.html

Design/Manufacturing for Hardware Security

- Arm race between attacking and protection
- Hardware IP reverse engineering using learning techniques
- Intelligent IC camouflaging [Li+, ICCAD'16, TCAD'17, HOST'17 BPA]
- **Former PhD Meng Li won ACM SRC Grand Finals First Place in 2018**



Outline

- Introduction
- AI for IC
- **IC for AI**
- Conclusion

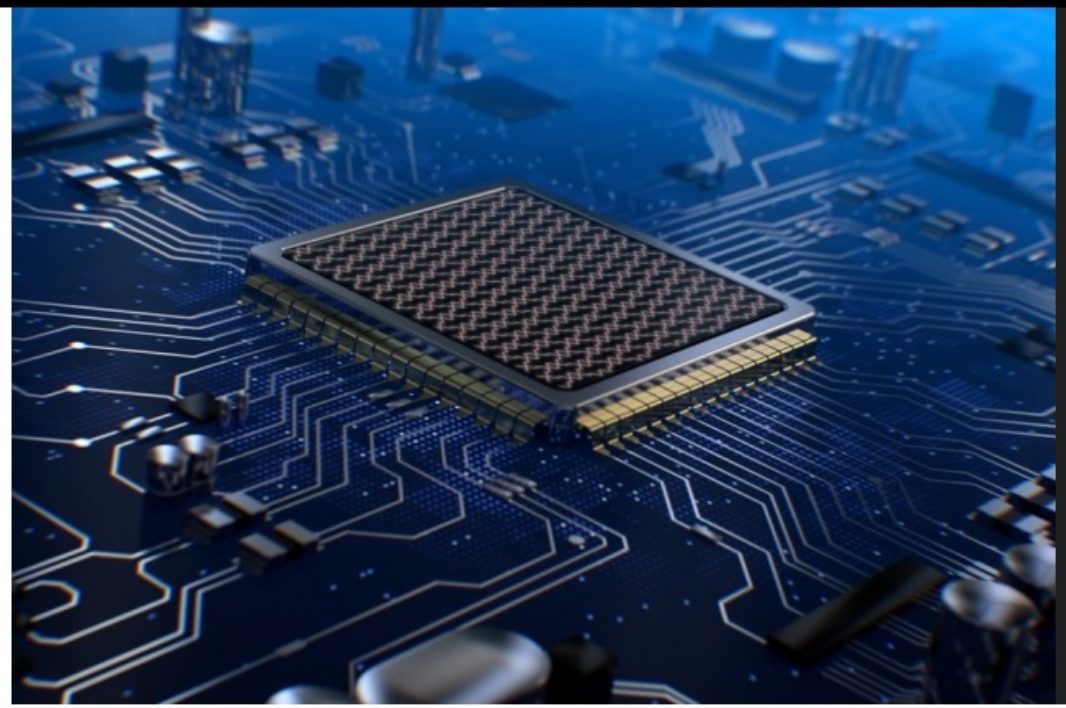
Photonic AI Chips

Based on optics/photonics ➔ photonic ICs

MIT News

ON CAMPUS AND AROUND THE WORLD

Browse or Search



FULL SCREEN

This futuristic drawing shows programmable nanophotonic processors integrated on a printed circuit board and carrying out deep learning computing.

Image: RedCube Inc., and courtesy of the researchers

New system allows optical “deep learning”

Neural networks could be implemented more quickly using new photonic technology.

LIGHTMATTER

OpenLight

LIGHTELLIGENCE

OPTELLIGENCE

Lighton

CogniFiber

XANADU

LUMINOUS

OPTIUS

light powered computing

SYNOPSYS PHOTONIC SOLUTIONS

cadence PHOTONICS

Ansys LUMERICAL

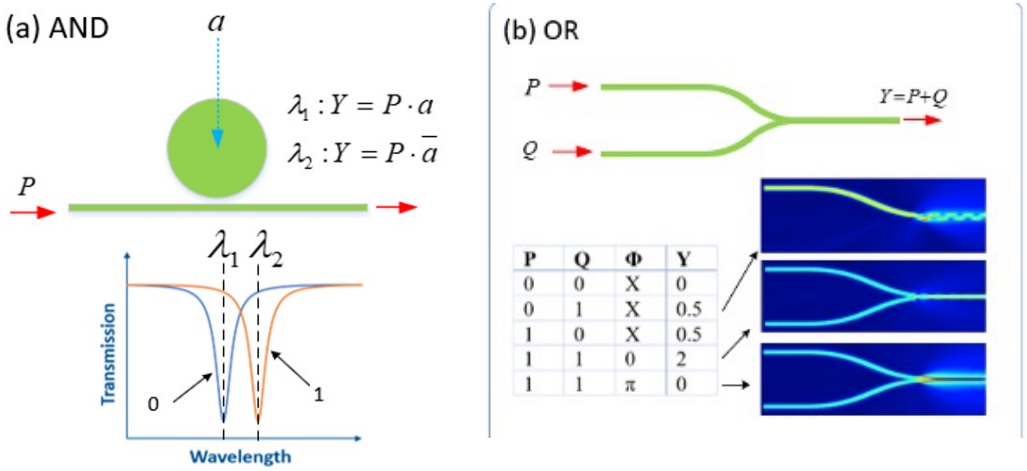
amf ADVANCED MICRO FOUNDRY

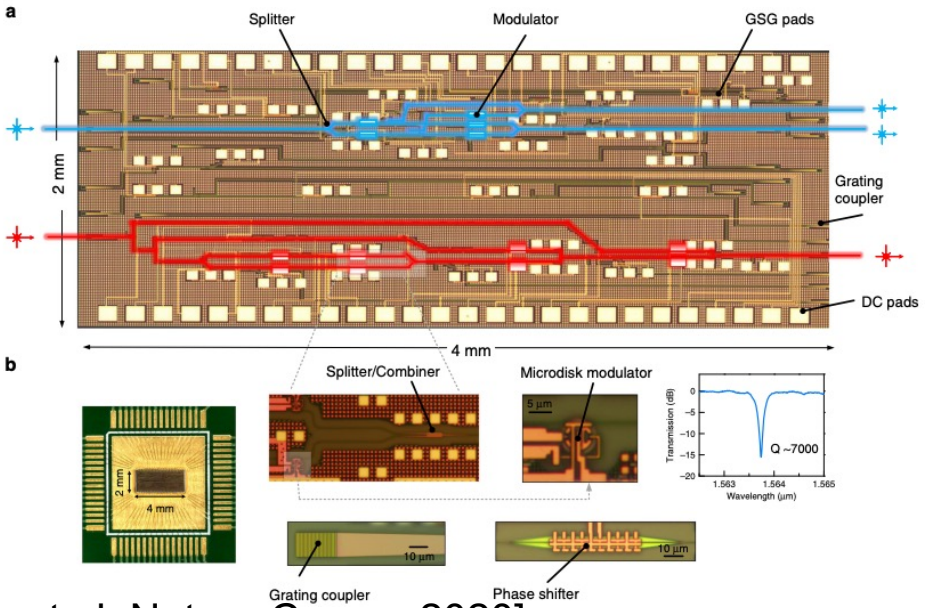
AIM photonics

GlobalFoundries

Optical Computing Basics

Digital computing

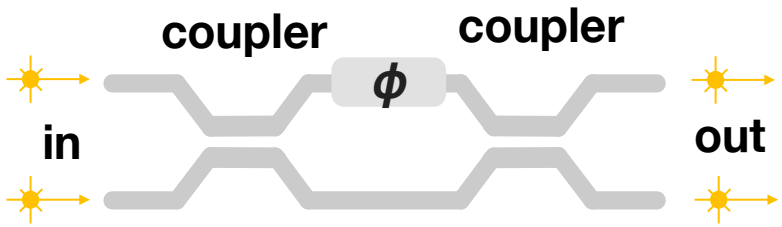




[Ying et al, Nature Comm. 2020]

Analog computing

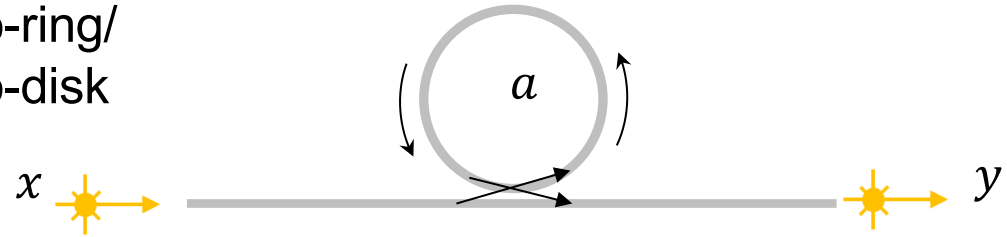
MZI



$$\begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$$

2×2 Unitary Matmul (~100×20 μm^2)

Micro-ring/
Micro-disk



$y = a \cdot x$

Scalar Mul. (~10×10 um^2)

WDM+PD

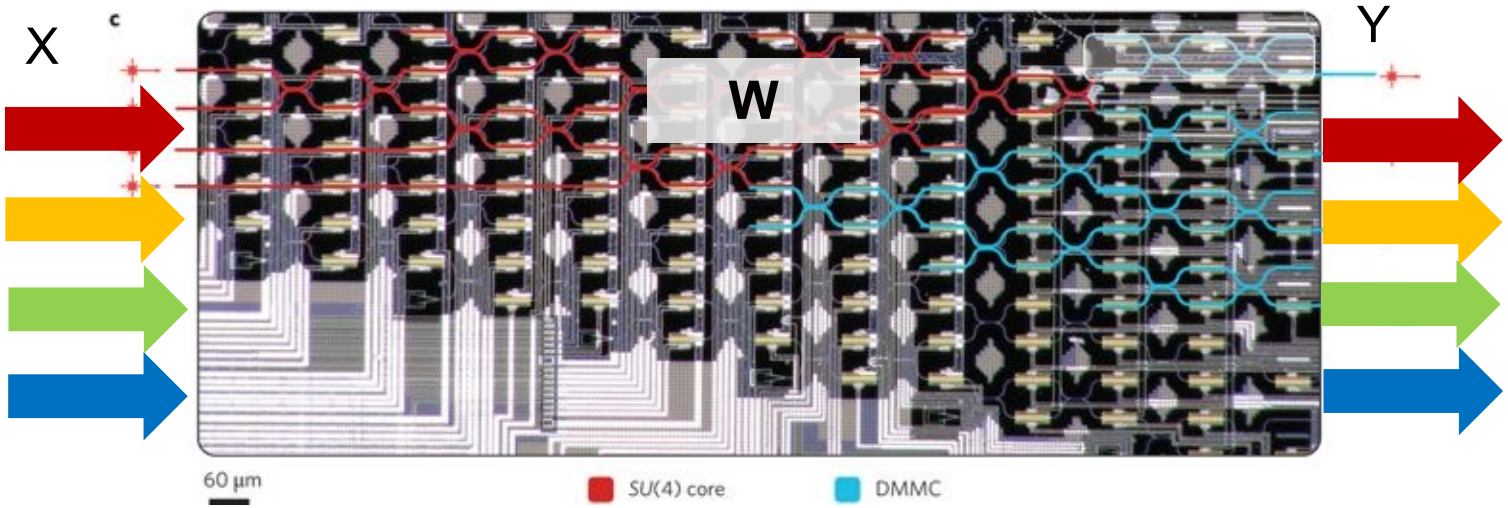
[Lightelligence]

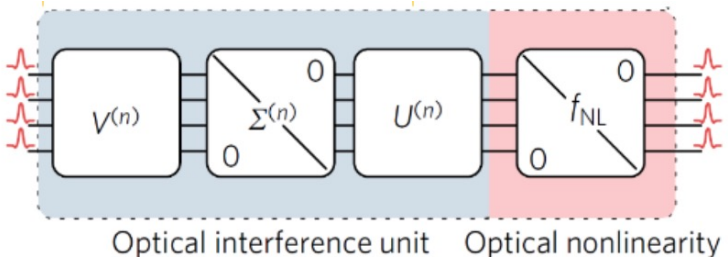
$y = \sum x_i$

WDM-based Summation

ONN Background: Photonics GEMM

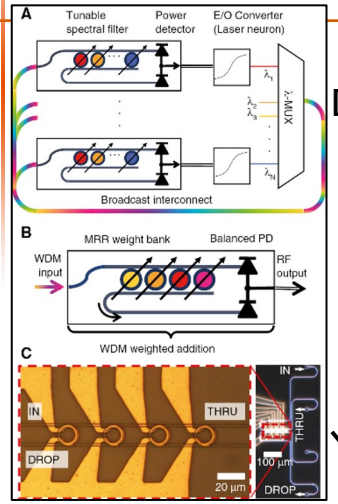
- DNNs: linear projection + nonlinear activation
  - Matrix multiplication is computation-intensive
- Photonics is good at ultra-fast linear operations



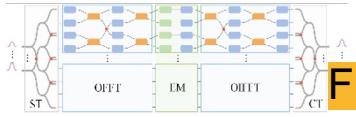


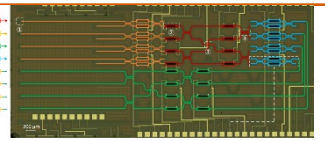
Photonic tensor unit for analog GEMM
[MIT's Nature Photonics'17]

ONN Ckt/Arch

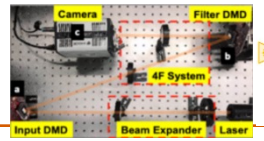


MRR ONN
[Brunner+, 2016]
[Tait+, SciRep 2017]
Princeton



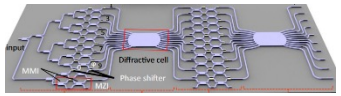


FFT-based optical neural network
[Gu+, ASPDAC2020]
UT

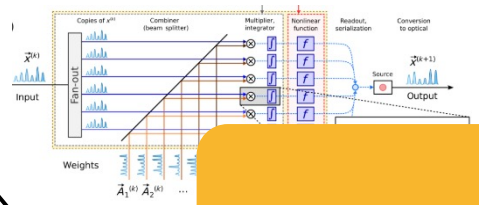


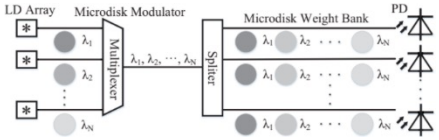
Free-space ONN
[Miscuglio+, Optica2020]
GWU

MORR ONN
[Gu+, DATE2021]
UT

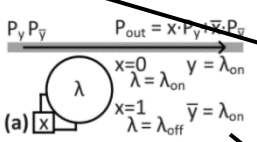


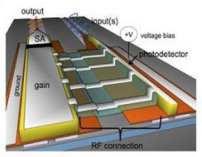
Star Coupler
[Zhu+, NatureComm2022]
NYU

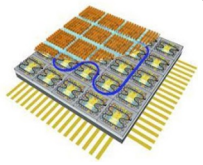




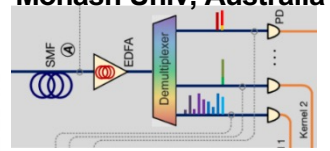
Holylight and Lightbulb:





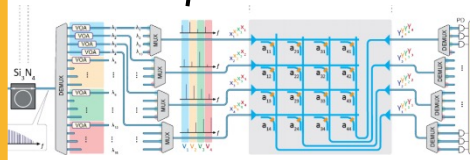


Optical Spike Neural Network
[Tait+, 2016]
Princeton



WDM Comb
[Xu+, Nature2021]
Monash Univ, Australia

PCM Xbar
[Miscuglio+, APR2020]
GWU



PCM Xbar
[Feldmann+, Nature2021]
Munster, Oxford

Area/Efficiency

Robustness

Learnability

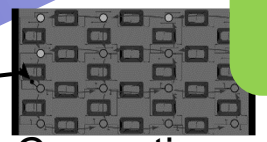
Circuit-Architecture-Algorithm Co-Design

2016

2020

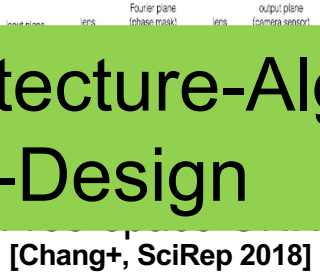
2021

2022

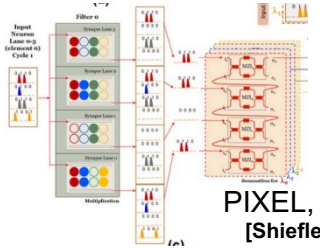


Optical Reservoir Computing
[Vandoorne+, NatureComm 2014]
Ghent University

MZI-based Neural Network
[Shen+, Nature Photonics 2017]
MIT

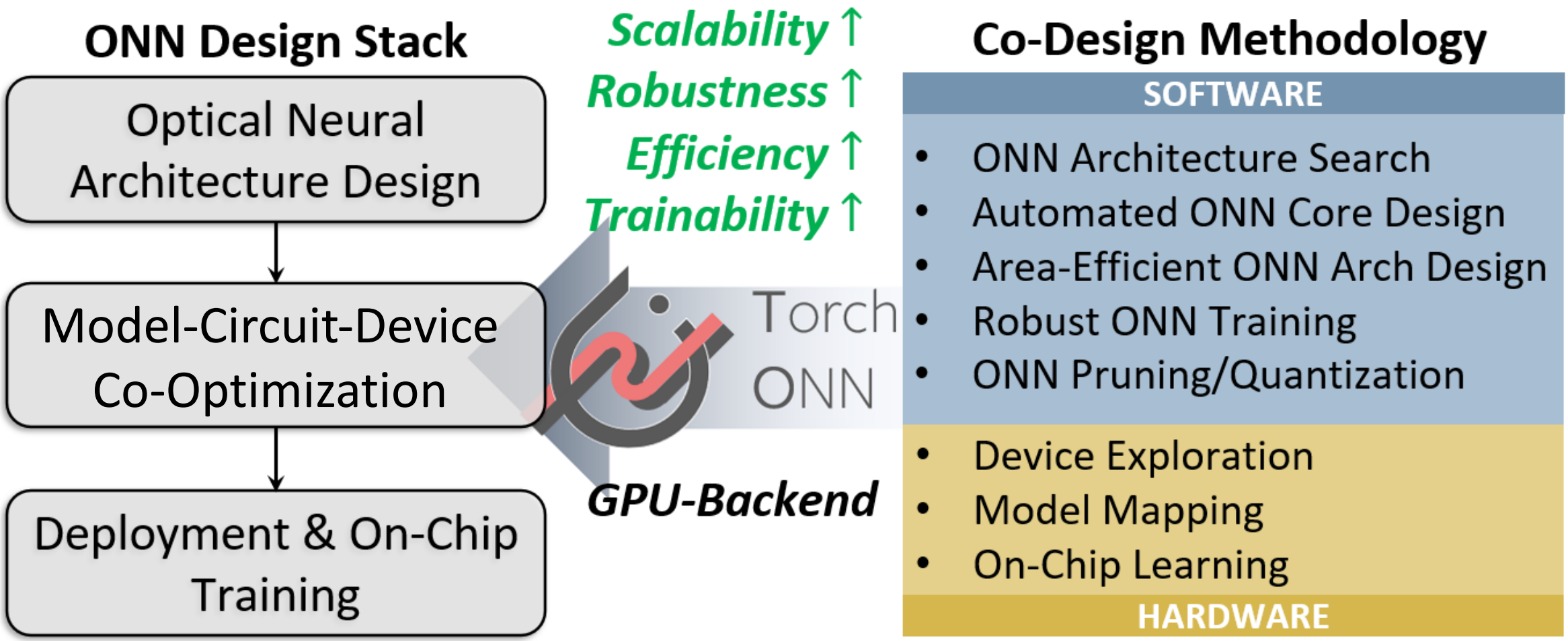


[Chang+, SciRep 2018]
Stanford



PIXEL, MZI Multiplier
[Shieflett+, HPCA2020]
Ohio Univ

Device-Circuit-Arch-Algorithm Co-Design Stack

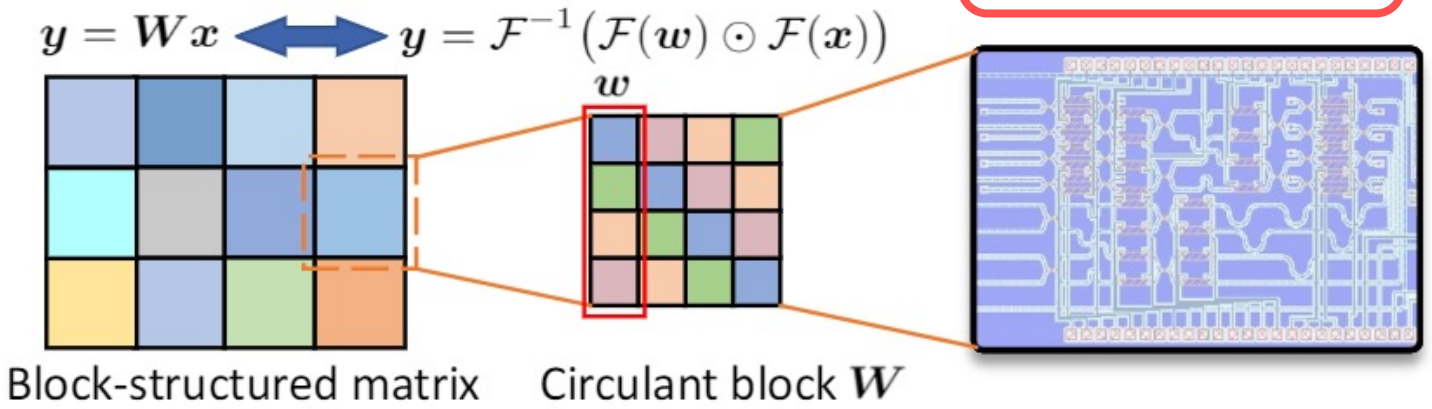


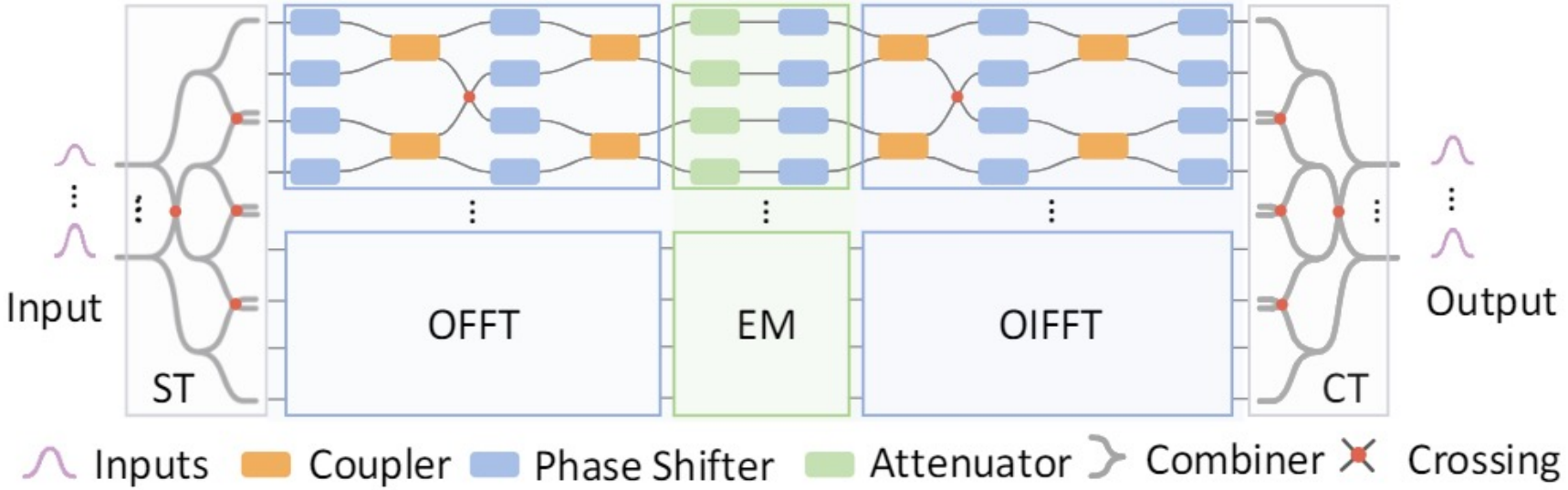
Jiaqi Gu won ACM Student Research Competition Grand Finals 1st Place 2021

Case Study 4 FFT-based ONN [Gu+, ASPDAC'20 BPA]

- Efficient circulant matrix multiplication in Fourier domain
- 2.2~3.7× area reduction, no accuracy loss

$$O(m^2 + n^2) \longrightarrow O\left(\frac{mn}{k}\log_2 k\right)$$

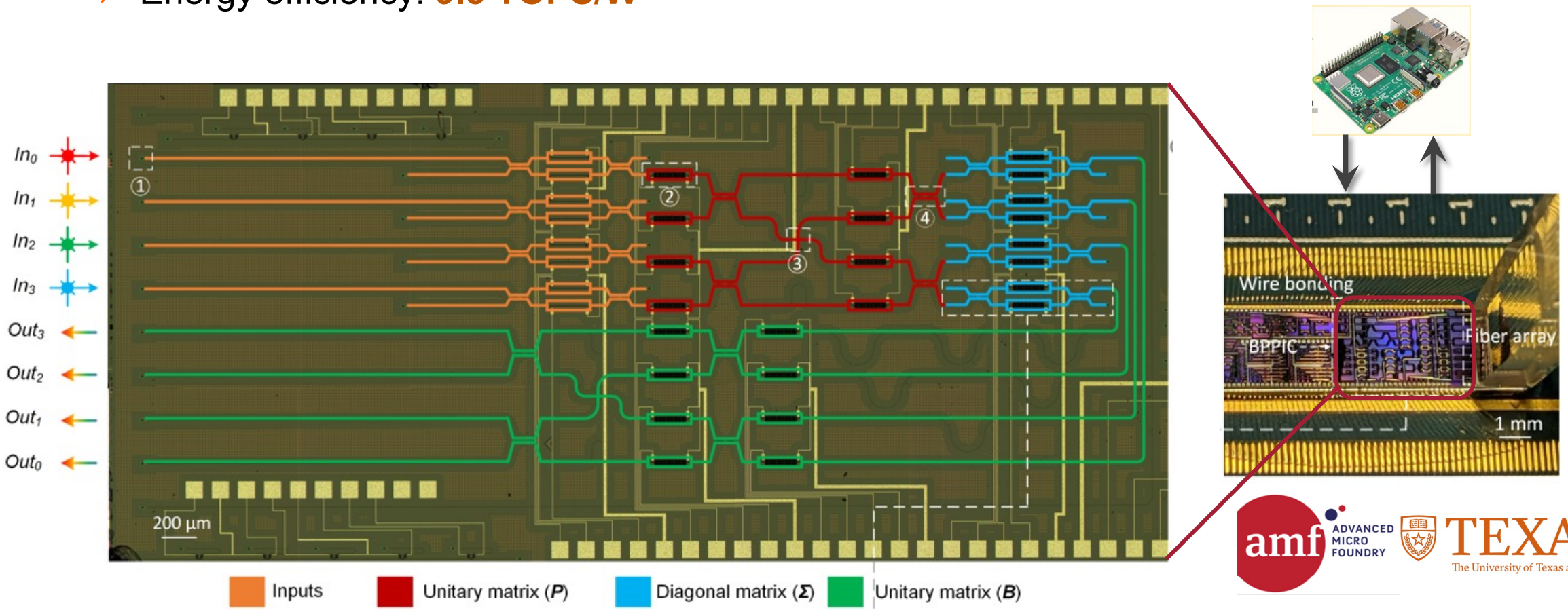




Our OSNN Neural Chip Tapeout & Measurement

- Experimental demonstration
  - Compute density: **225 TOPS/mm²**
  - Energy efficiency: **9.5 TOPS/W**

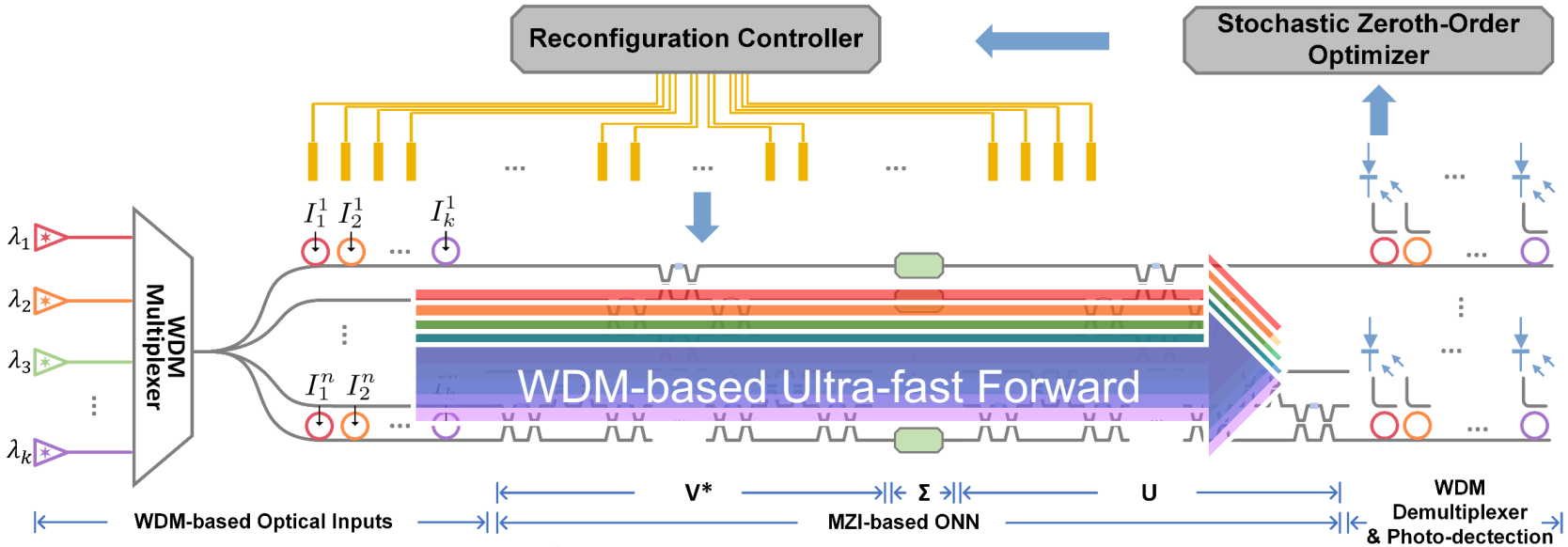
Won the Robert S. Hilbert Memorial Optical Design Competition, July 2022

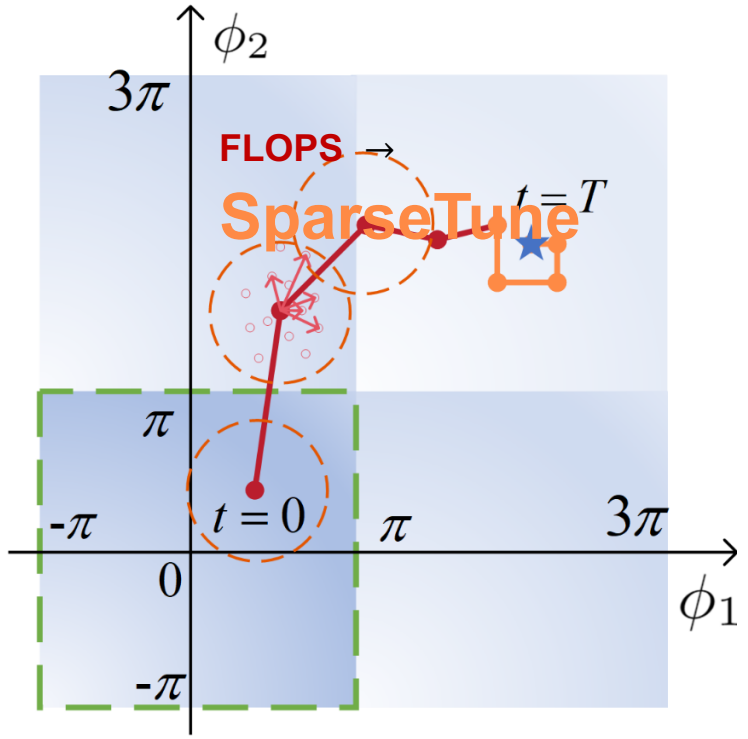


C. Feng, J. Gu, H. Zhu, Z. Ying, Z. Zhao, D.Z. Pan, R.T. Chen, Under Submission

Case Study 5 FLOPS [DAC'20 BPC] [NSF Workshop'20, BPA]

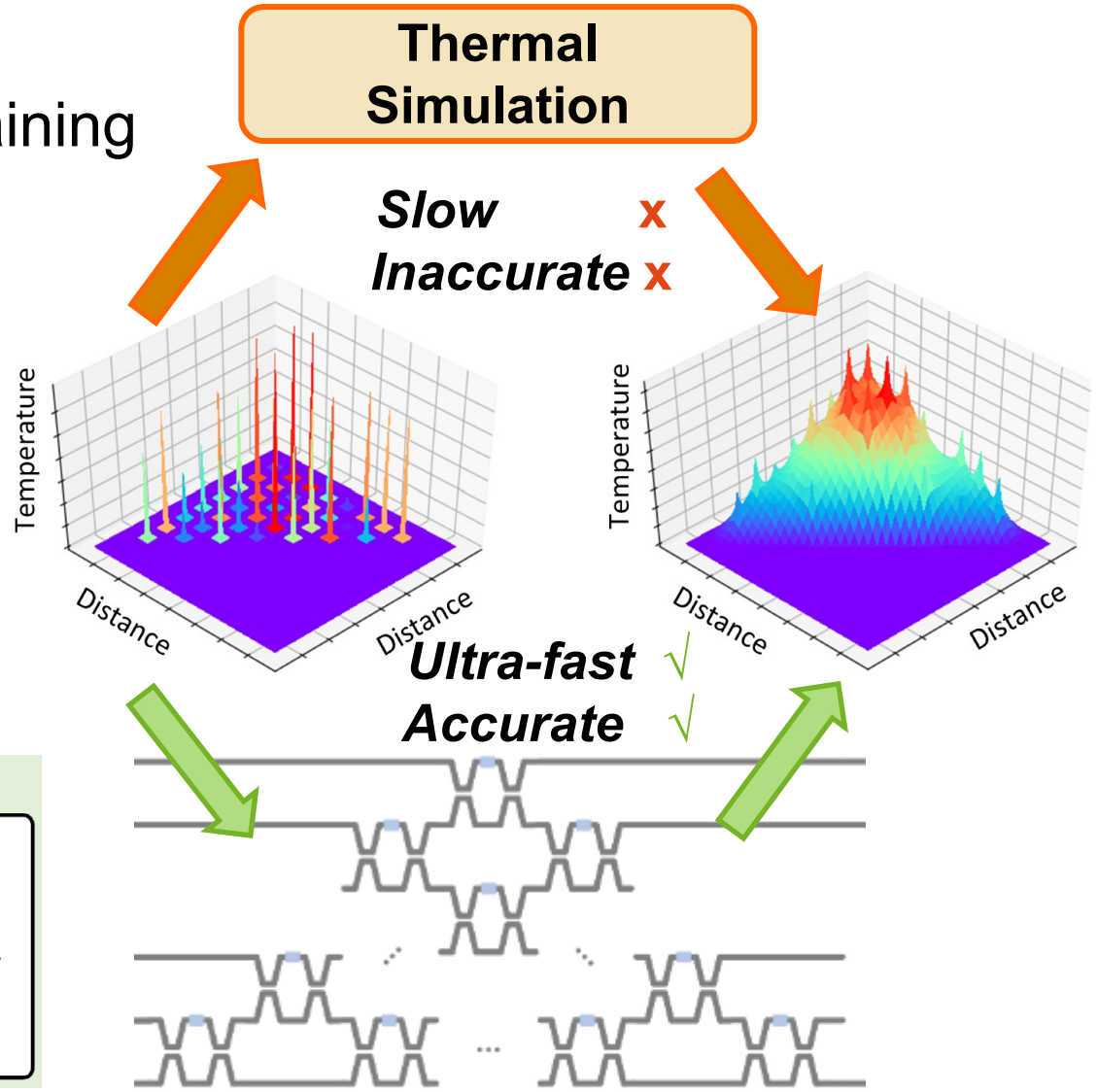
- ONN on-chip learning via stochastic zeroth-order optimization
  - **Efficiency**: WDM-based forward-only gradient estimation
  - **Accuracy**: Two-stage learning protocol (FLOPS+) with high accuracy
  - **Robustness**: Robust learning under *in situ* device variations

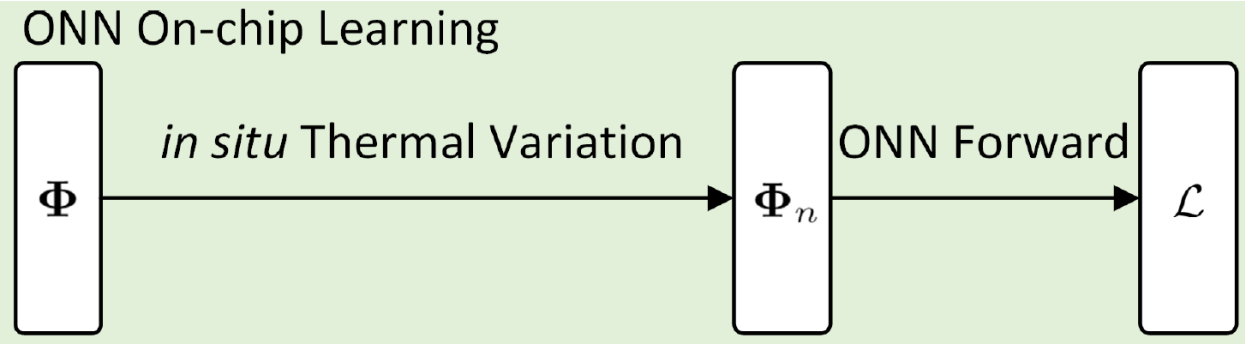




Robust On-Chip Learning

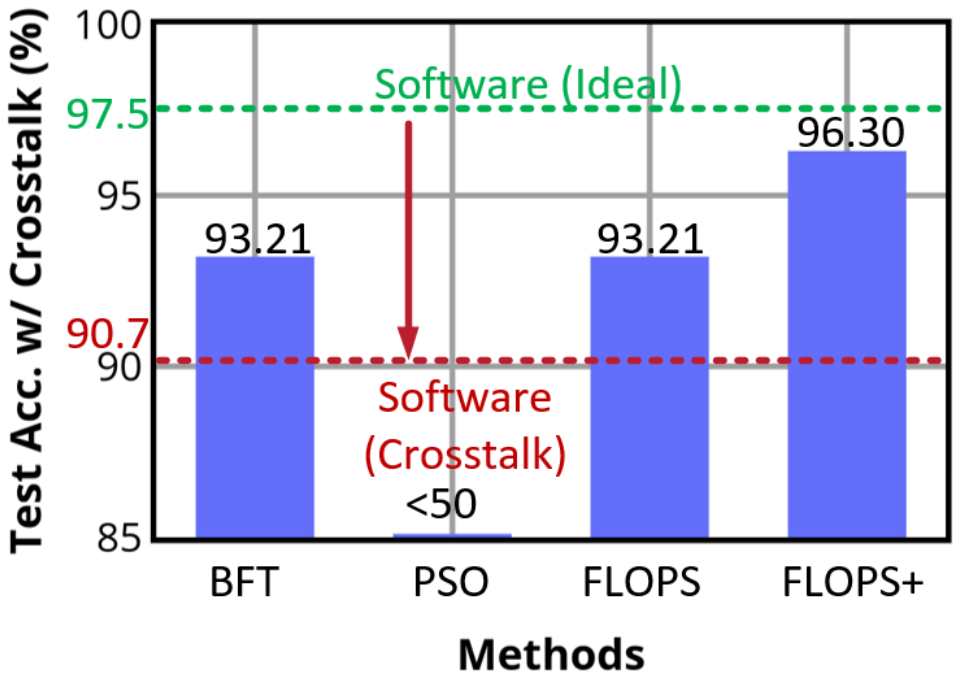
- Thermal crosstalk variations
  - Typically not considered in software training
  - Time-consuming
  - Inaccurate
- Built-in robustness handling on-chip
  - Ultra-fast: ~1 μs
  - Accurate: physical noise model





Experimental Results [Gu+, DAC'20]

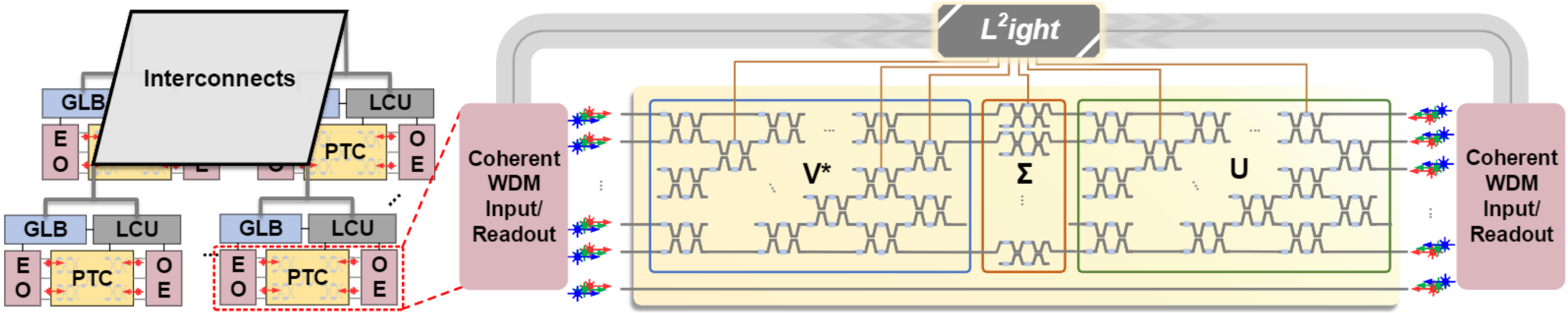
- Robust learning under *in situ* thermal variations
  - **5%** more accurate than hardware-agnostic software training
  - **3%** more robust than previous on-chip training approaches



ONN config: 10-24-24-6 (960 MZIs)

L²ight – Scalable On-Chip Training [Gu+, NeurIPS'21]

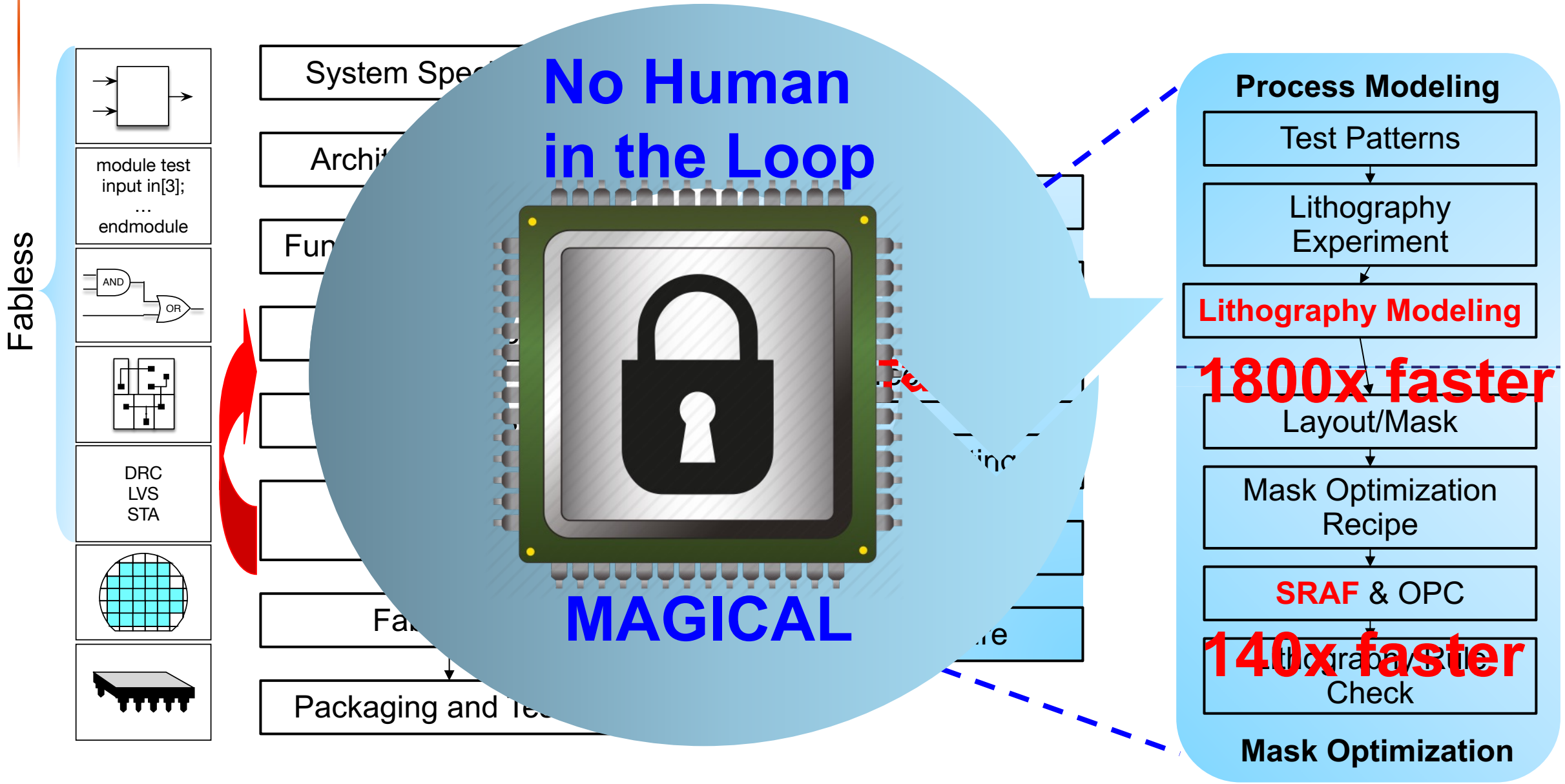
- Gradient-free methods ➔ **First-order gradient-based**
- Can handles **million**-parameter ONNs
  - **1000×** more scalable than [Gu+, DAC'20] to handle million-parameter ONNs
  - Efficiency: Multi-level sparsity to boost efficiency by ***30×***
- In-situ noise consideration for *noise-resilient* ONNs



Outline

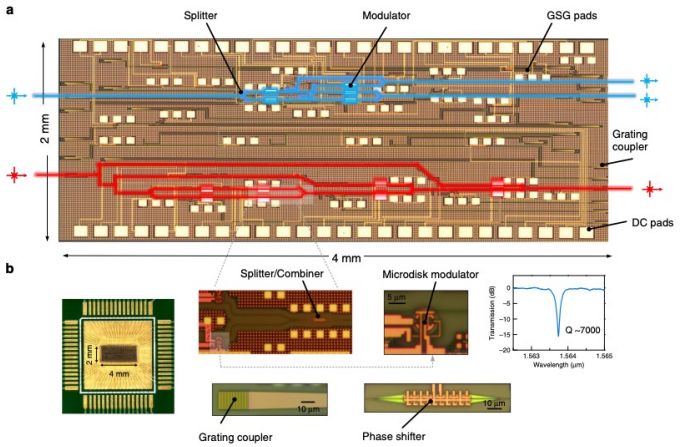
- Introduction
- AI for IC
- IC for AI
- Conclusion

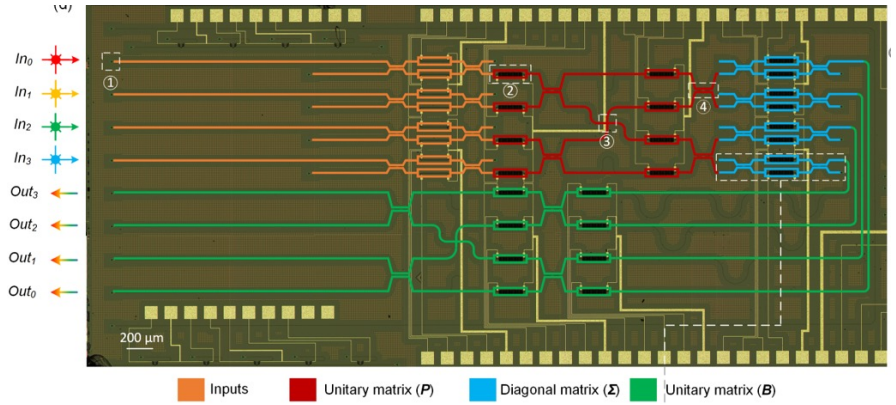
To Recap: AI for IC

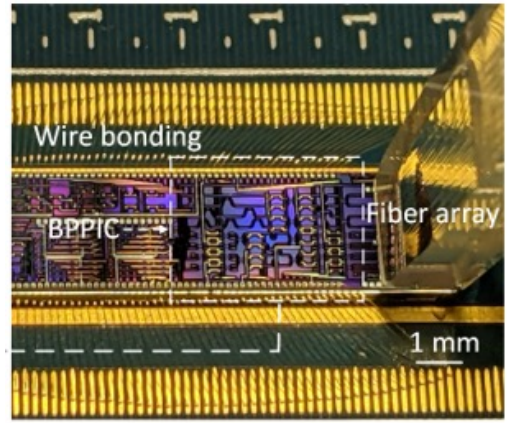


To Recap: [Photonic] IC for AI

- How to build ultra-fast (light-speed) and ultra-efficient optical neural accelerators with photonic integrated circuits
  - Software and hardware co-design is **KEY**
- FFT-ONN (**ASP-DAC 2020 Best Paper Award**)
- FLOPS (**DAC 2020 Best Paper Finalists; NSF'20 Workshop BPA**)
- PhD student Jiaqi Gu won **ACM SRC Grand Finals 1st Place** in 2021
- **Robert S. Hilbert Memorial Optical Design Competition**, July 2022

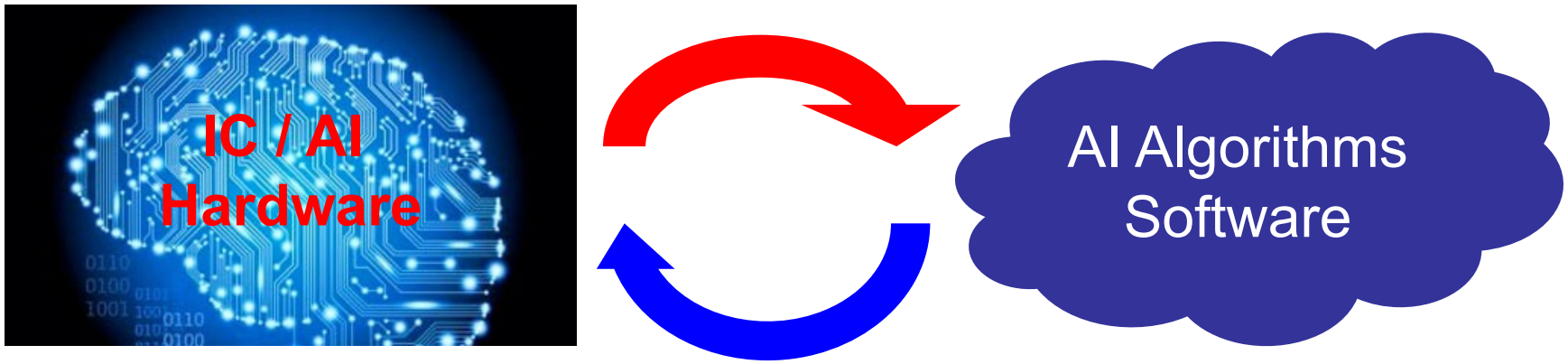






Conclusion

- Advance in AI algorithms/software ➔ Agile IC/hardware design
- Advance in IC/hardware ➔ Enhanced AI capability



Closing the Virtuous Cycle!

Acknowledgment

- Funding support / collaborations from NSF, DARPA, MURI, Intel, Nvidia, Google, Synopsys, Toshiba Memory (Kioxia), AMD/Xilinx, VMware, etc.
- **Many students/post-docs** who do the real work
- Many collaborators
  - Dr. Haoxing Ren from NVIDIA for DREAMPlace, AutoCRAFT
  - Prof. Nan Sun at UT Austin (now at Tsinghua) for MAGICAL
  - Dr. Nojima et al. from Toshiba Memory (KIOXIA) on DFM
  - Prof. Ray Chen at UT Austin for optical interconnect/computing
  - …

Thanks!

Q & A?