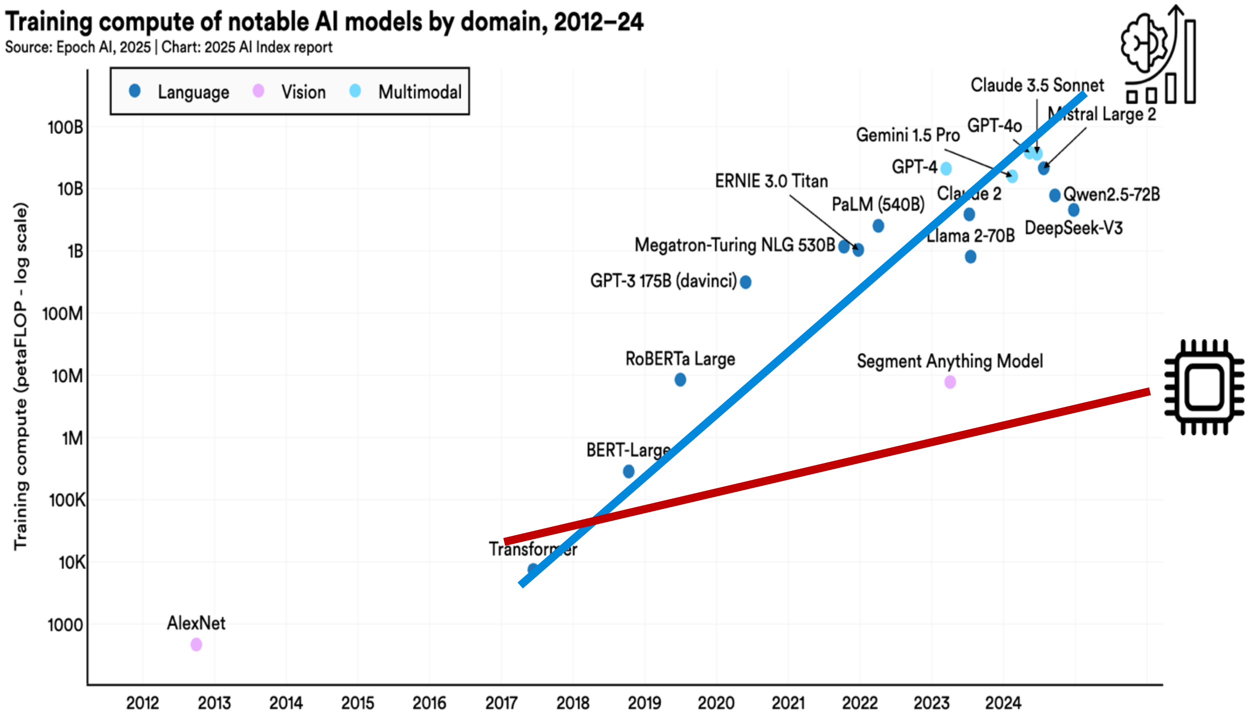
DAC DCgAA Workshop Keynote, June 22, 2025

# Closing the Generative AI–Hardware Loop: Photonic Acceleration, Memory-Efficient Training, and AI-Driven IC Design

David Z. Pan
Electrical & Computer Engineering
The University of Texas at Austin

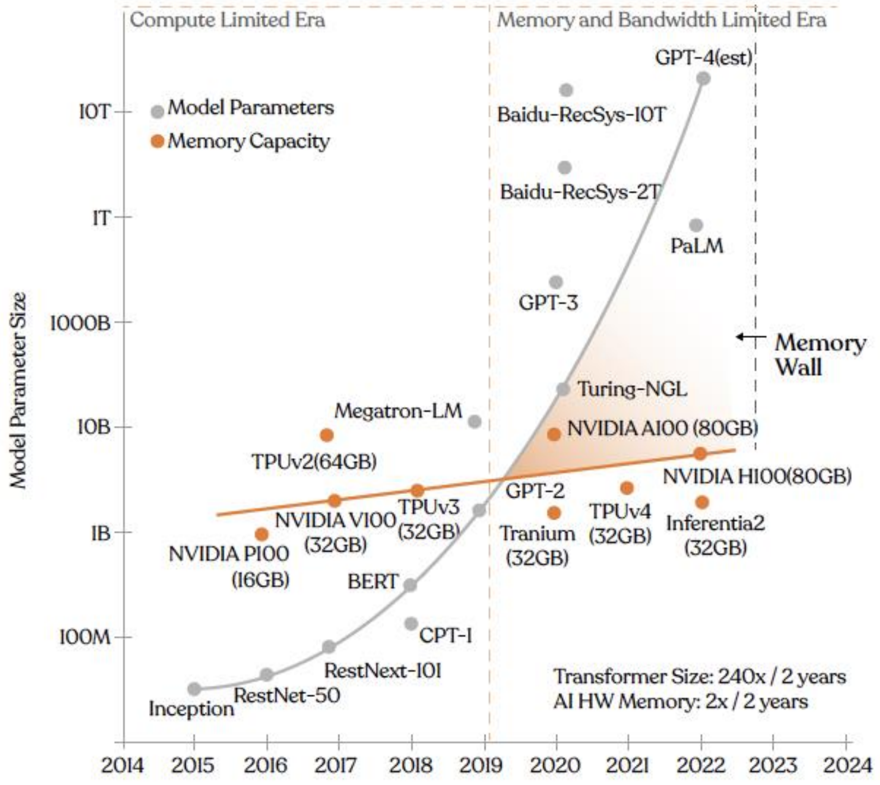
# AI Model Scaling Hits Hardware Wall

- AI compute/model scaling doubles every 5 months
- Significantly outpace **Moore’s** law and **Memory scaling**


Training compute of notable AI models by domain, 2012–24

Source: Epoch AI, 2025 | Chart: 2025 AI Index report

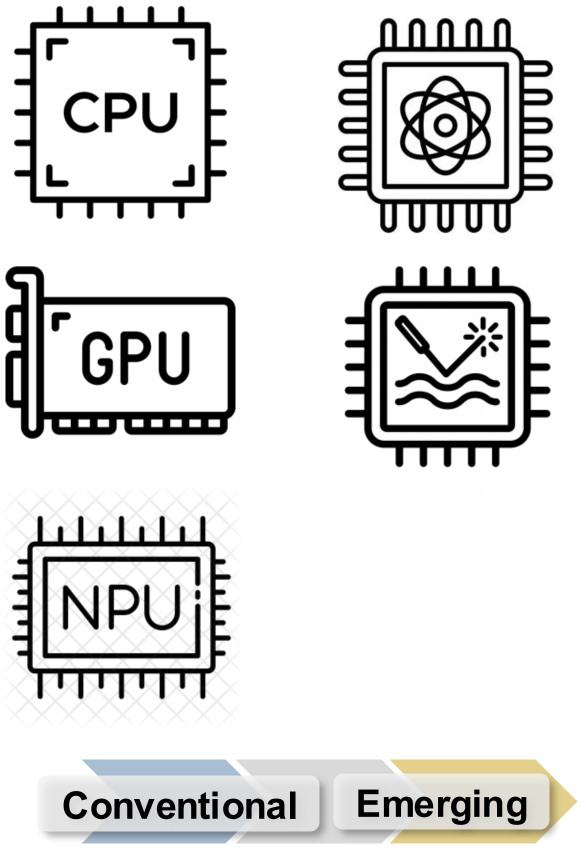
Compute scaling



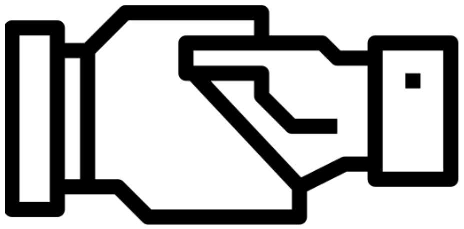
Memory Capacity scaling

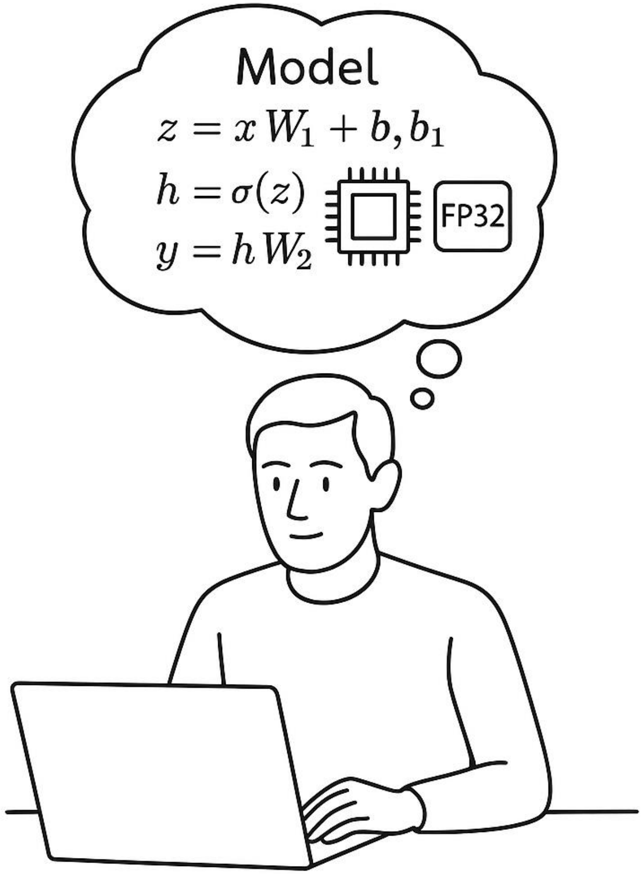
# AI Model and Hardware Co-Design

- Traditionally, fixed ML algorithms → then systems optimization
- Invent ML algorithms that are **hardware-aware** and ML algorithm, system, hardware co-design

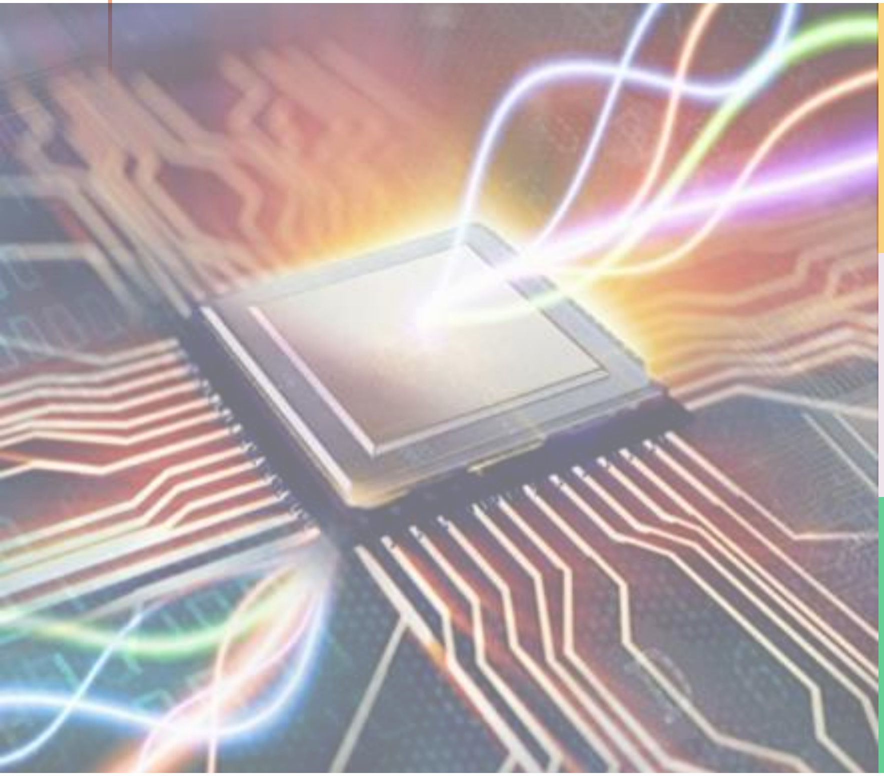


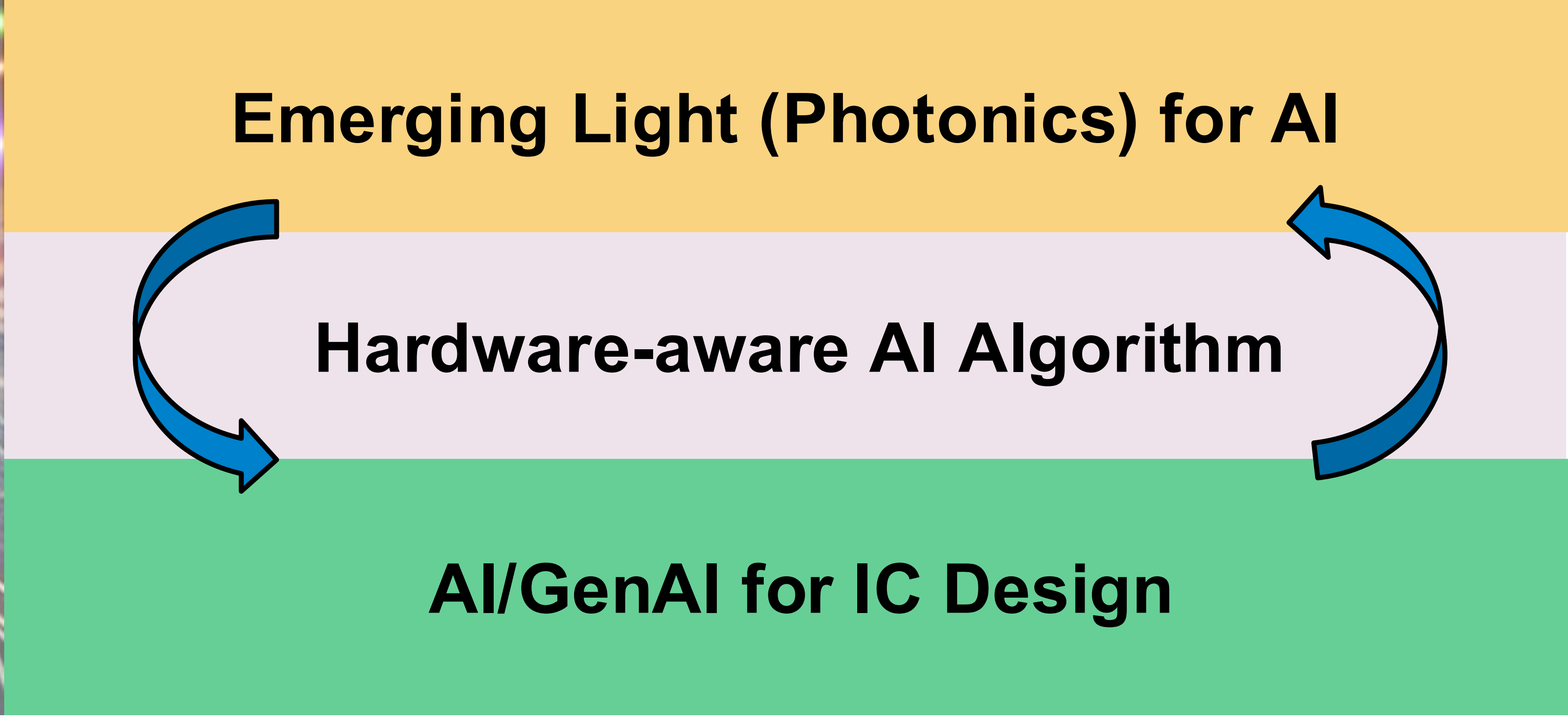
Scalable
Fast
Accurate
…





# In This Talk: AI-Hardware Synergy



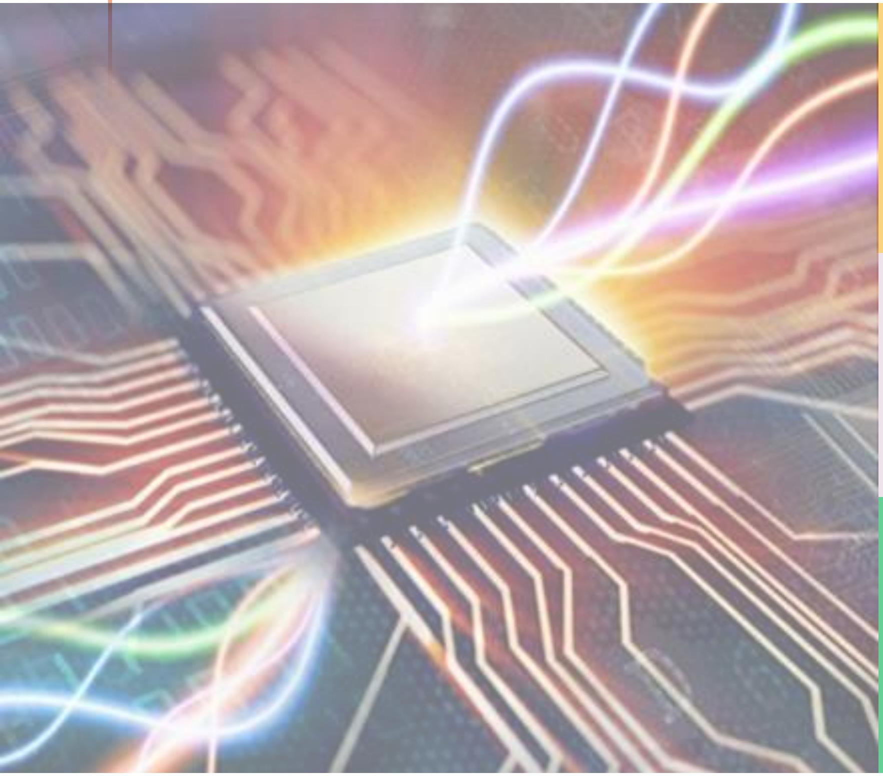


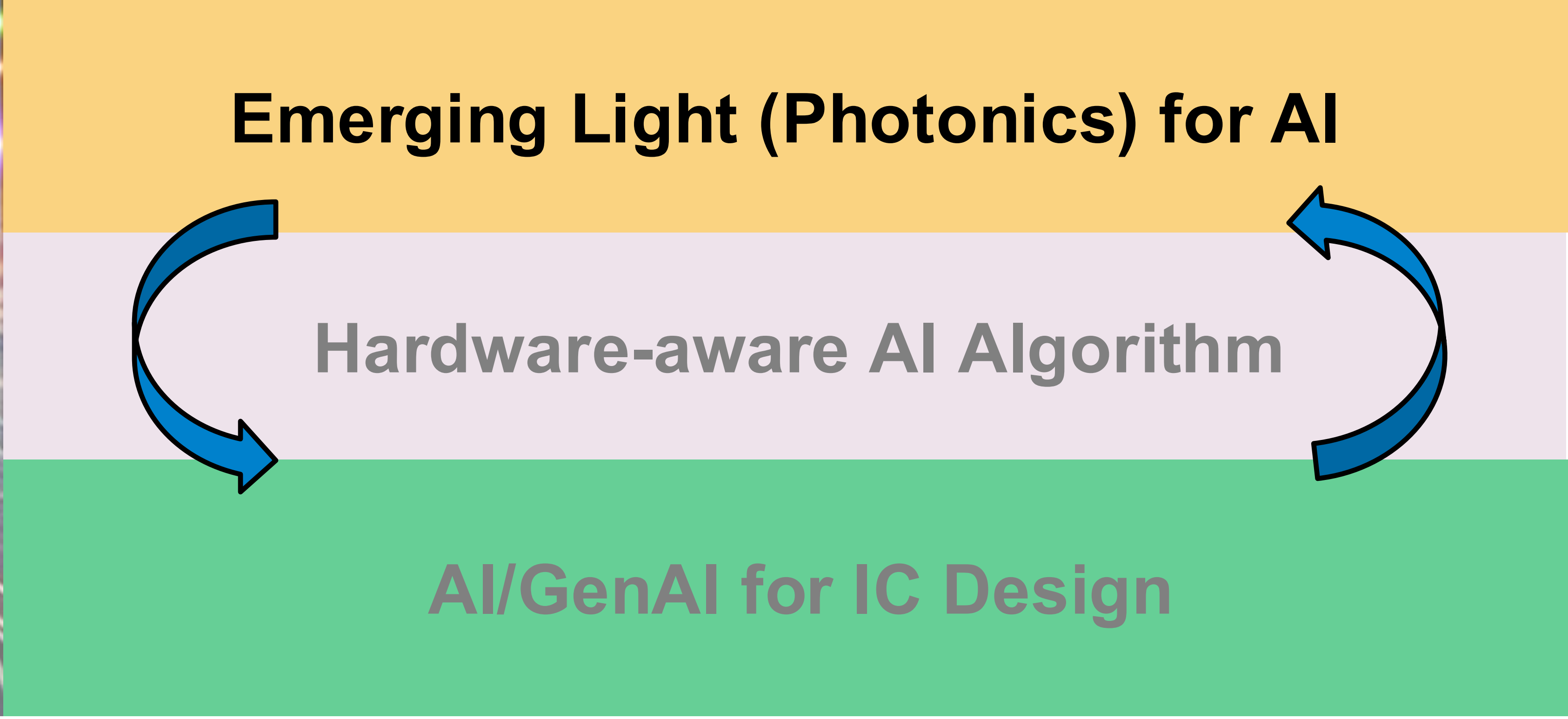
Emerging Light (Photonics) for AI

Hardware-aware AI Algorithm

AI/GenAI for IC Design

# In This Talk: AI-Hardware Synergy





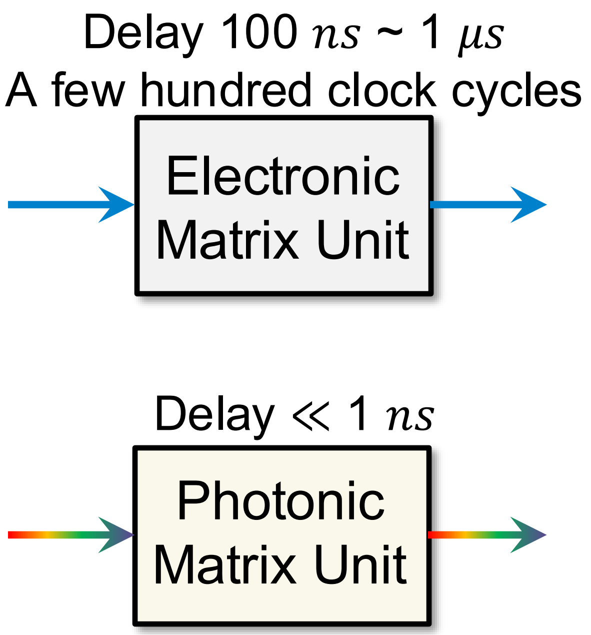
Emerging Light (Photonics) for AI

Hardware-aware AI Algorithm

AI/GenAI for IC Design

# Electrical Computing vs Photonic Computing

## High speed



Delay 100 $ns$ ~ 1 $\mu s$

A few hundred clock cycles

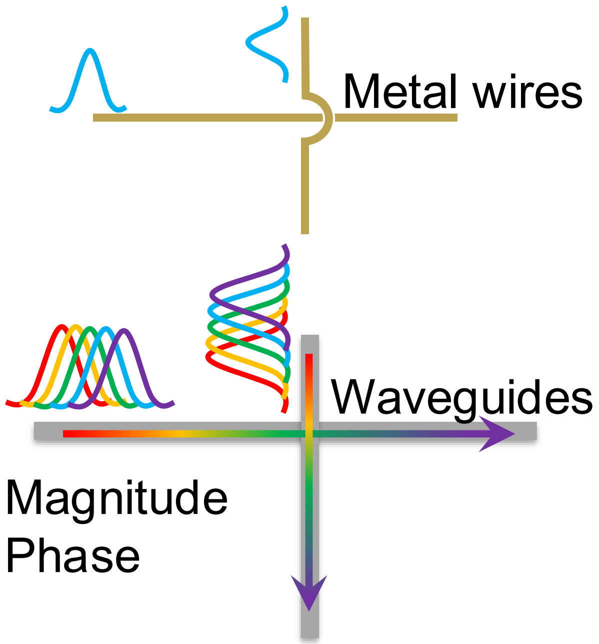
Electronic Matrix Unit

Delay ≪ 1 $ns$

Photonic Matrix Unit

Computing as light propagate

## Massive parallelism



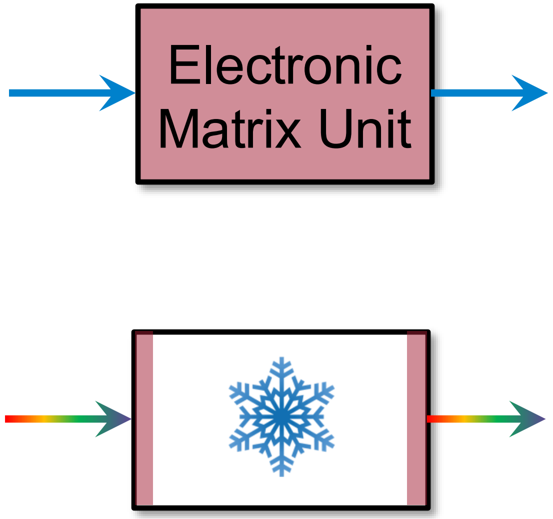
Metal wires

Waveguides

Magnitude
Phase

Light propagate in parallel

## High energy efficiency



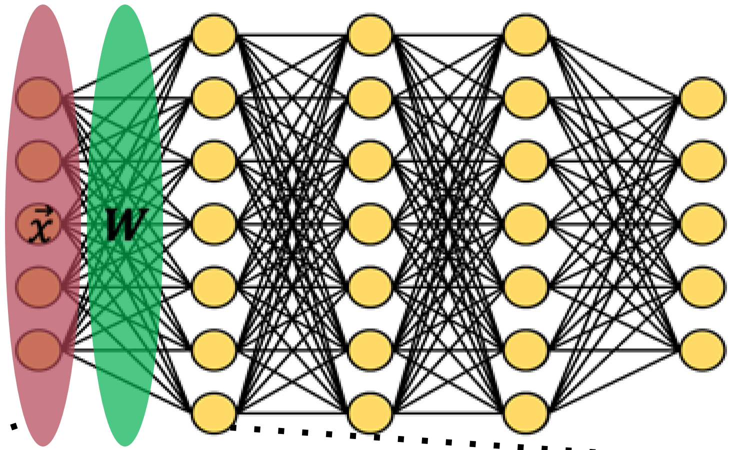
Electronic Matrix Unit

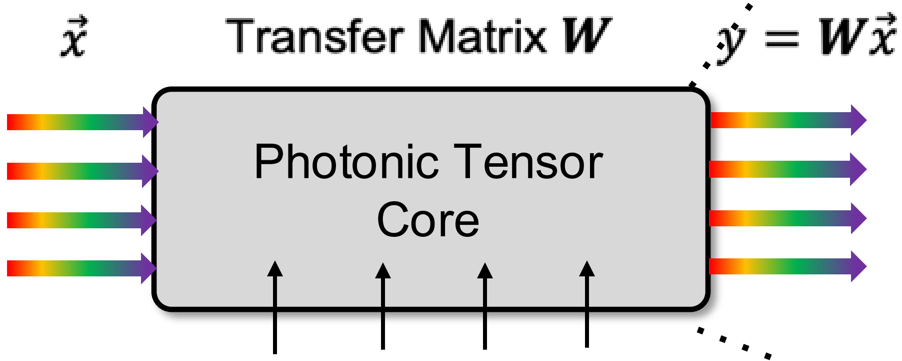
Passive circuits consumes near zero static power

Source: https://www.lightelligence.ai/

# General Photonic AI Computing Paradigm

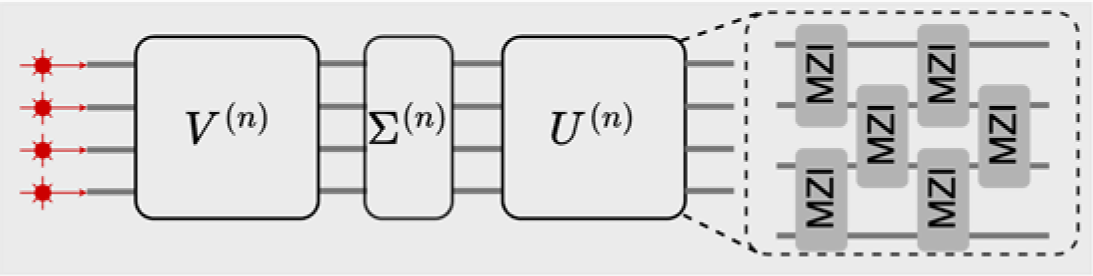
- Encode weight matrix into photonic circuit transformation
- Efficient one-shot $\boldsymbol{W}x$ by forwarding optical signal





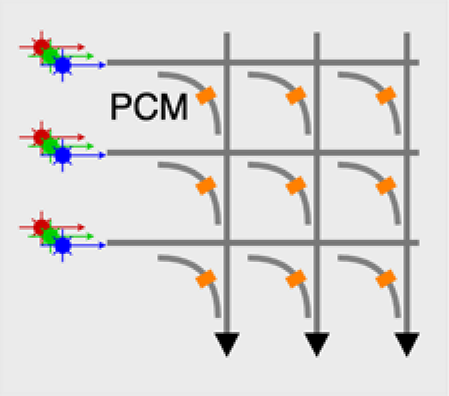
$\vec{x}$
Transfer Matrix $\boldsymbol{W}$
$\vec{y} = \boldsymbol{W}\vec{x}$
Photonic Tensor Core
Optical Inputs
Electrical Weight Encoding
Optical Outputs

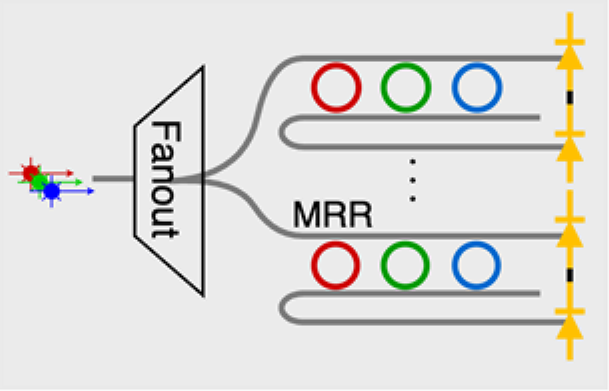
**Representative Photonic Tensor Cores**



MZI Mesh

[Nat. Photon'17] [Science'23]





PCM Crossbar

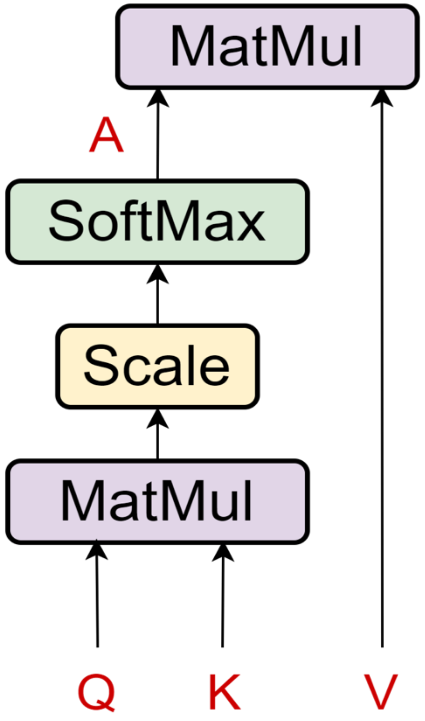
[Nature'21]

MRR Bank

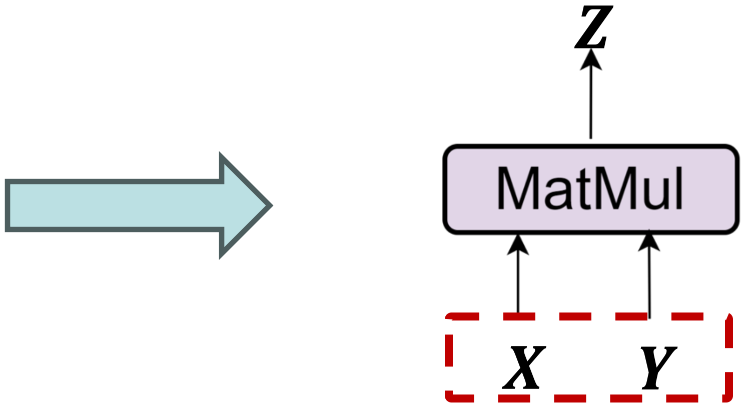
[SciRep'17]

# Move to Support GenAI Workloads: Transformers!

- ***Prior Optic AI → Designed for CNN (fixed weights and positive Inputs)***
- **Dynamic** Matrix Multiplication
  - ***Real-time** operand programming*
- **Full-range** Operands

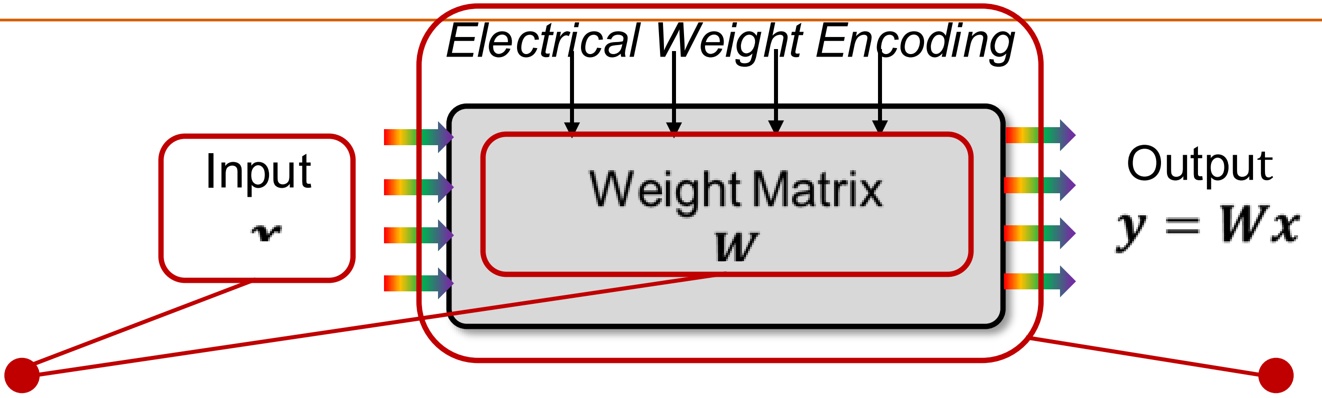


Attention



***Dynamical*** and ***full-range*** activations

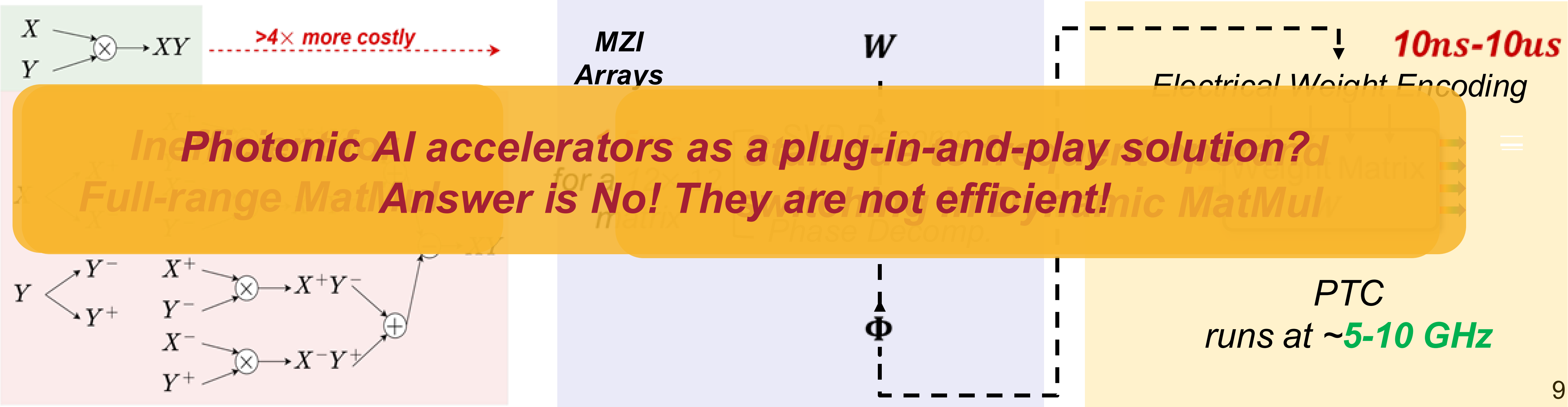
# Prior PTCs as Plug-in-and-Play Solutions? No!



**Non-negative only operands in incoherent PTC** (light intensity modulation)

**High operand mapping Cost** (time-consuming decomposition step)

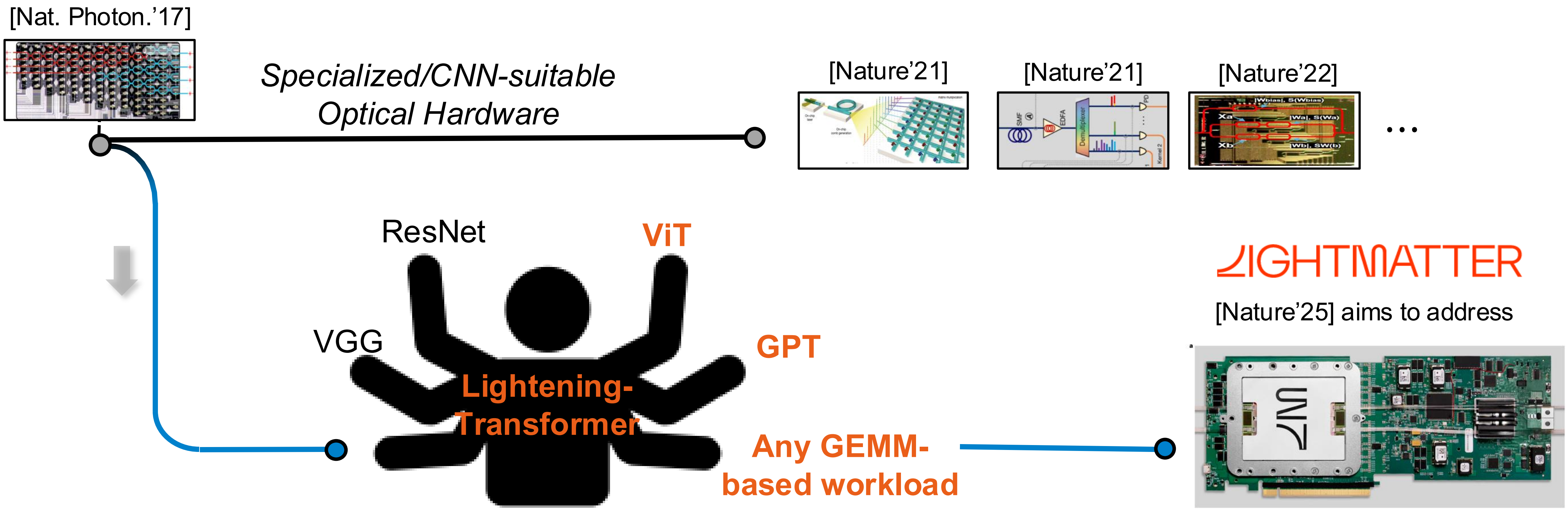
**Slow device programming** (adoption of low-loss, compact weight modulators)



***Photonic AI accelerators as a plug-in-and-play solution? Answer is No! They are not efficient!***

# Our Lightening-Transformer, HPCA'24

- The ***first* versatile** photonic accelerator for universal AI models
- Deliver **100-1000x** performance gain than electronics on Transformers
- First open-sourced arch-level simulator for optical AI accelerator



[Nat. Photon.'17]

*Specialized/CNN-suitable Optical Hardware*

[Nature'21] [Nature'21] [Nature'22] …

ResNet

VGG

**ViT**

**GPT**

**Lightening-Transformer**

**Any GEMM-based workload**

LIGHTMATTER

[Nature'25] aims to address

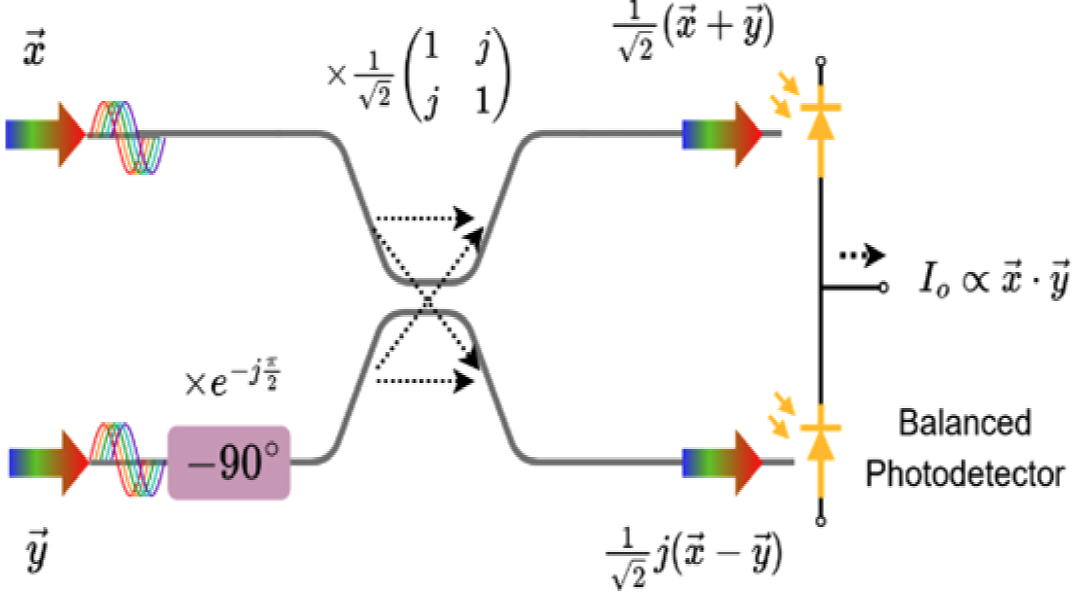
**Zhu, Hanqing**, Jiaqi Gu, Hanrui Wang, Zixuan Jiang, Zhekai Zhang, Rongxing Tang, Chenghao Feng, Song Han, Ray T. Chen, and David Z. Pan. "Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator.". **HPCA** 2024.

# A Novel Versatile Photonic GEMM Primitive

- Compute via ***direct light-light interaction***

High-speed encoded & Full-range optical inputs

*Fully-passive internal optical circuit*


$\vec{x}$
$\times\frac{1}{\sqrt{2}}\begin{pmatrix}1 & j\\ j & 1\end{pmatrix}$
$\frac{1}{\sqrt{2}}(\vec{x}+\vec{y})$
$I_o \propto \vec{x}\cdot\vec{y}$
$\times e^{-j\frac{\pi}{2}}$
$-90^\circ$
Balanced Photodetector
$\vec{y}$
$\frac{1}{\sqrt{2}}j(\vec{x}-\vec{y})$

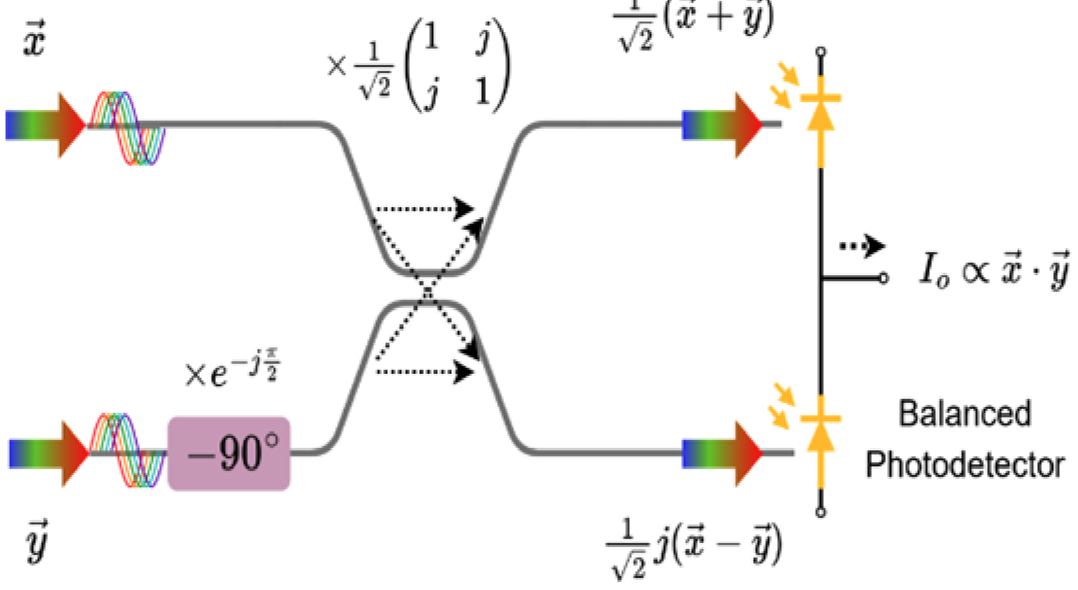
Dynamic dot-product engine: DDot

# A Novel Versatile Photonic GEMM Primitive

- Compute via ***direct light-light interaction*** → Dynamical modulation cost concern
- ***Crossbar-style*** photonic tensor core via optical broadcast
  - *Maximized intra-core operand sharing* for both X and Y (memristor crossbar: W(X) only)

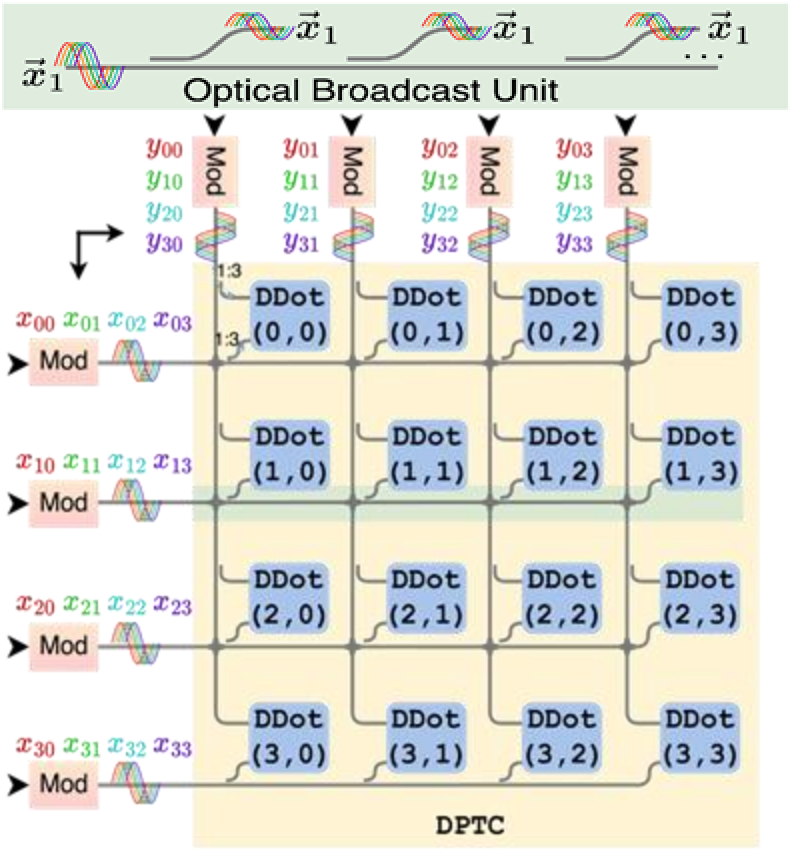
High-speed encoded & Full-range optical inputs

*Fully-passive internal optical circuit*



$\vec{x}$
$\times\frac{1}{\sqrt{2}}\begin{pmatrix}1 & j\\ j & 1\end{pmatrix}$
$\frac{1}{\sqrt{2}}(\vec{x}+\vec{y})$
$I_o \propto \vec{x}\cdot\vec{y}$
$\times e^{-j\frac{\pi}{2}}$
$-90°$
Balanced Photodetector
$\vec{y}$
$\frac{1}{\sqrt{2}}j(\vec{x}-\vec{y})$

Dynamic dot-product engine: DDot

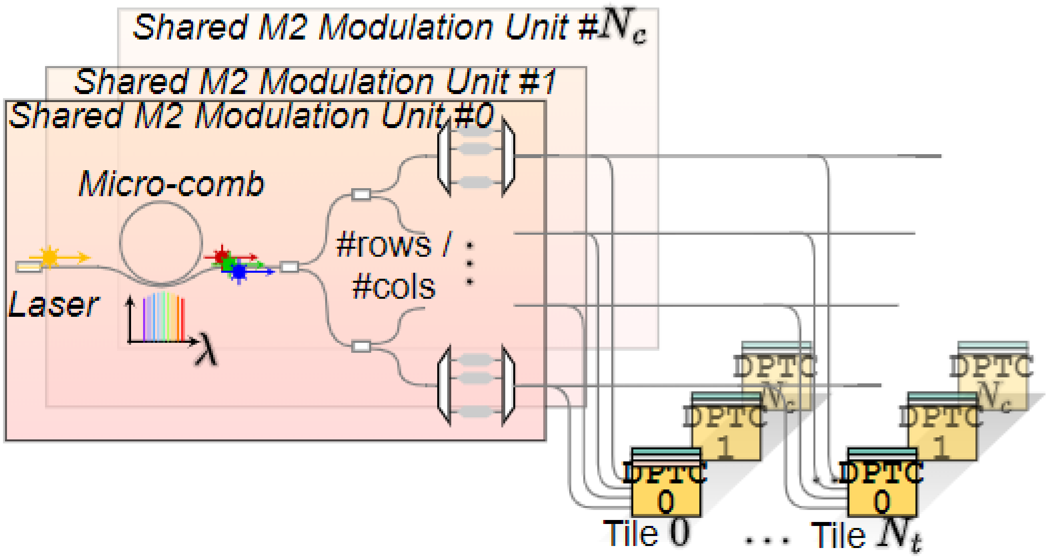


$\vec{x}_1$
Optical Broadcast Unit
Mod
$y_{00}$ $y_{10}$ $y_{20}$ $y_{30}$
$y_{01}$ $y_{11}$ $y_{21}$ $y_{31}$
$y_{02}$ $y_{12}$ $y_{22}$ $y_{32}$
$y_{03}$ $y_{13}$ $y_{23}$ $y_{33}$
1:3
$x_{00}$ $x_{01}$ $x_{02}$ $x_{03}$
$x_{10}$ $x_{11}$ $x_{12}$ $x_{13}$
$x_{20}$ $x_{21}$ $x_{22}$ $x_{23}$
$x_{30}$ $x_{31}$ $x_{32}$ $x_{33}$
DDot (0,0)
DDot (0,1)
DDot (0,2)
DDot (0,3)
DDot (1,0)
DDot (1,1)
DDot (1,2)
DDot (1,3)
DDot (2,0)
DDot (2,1)
DDot (2,2)
DDot (2,3)
DDot (3,0)
DDot (3,1)
DDot (3,2)
DDot (3,3)
DPTC

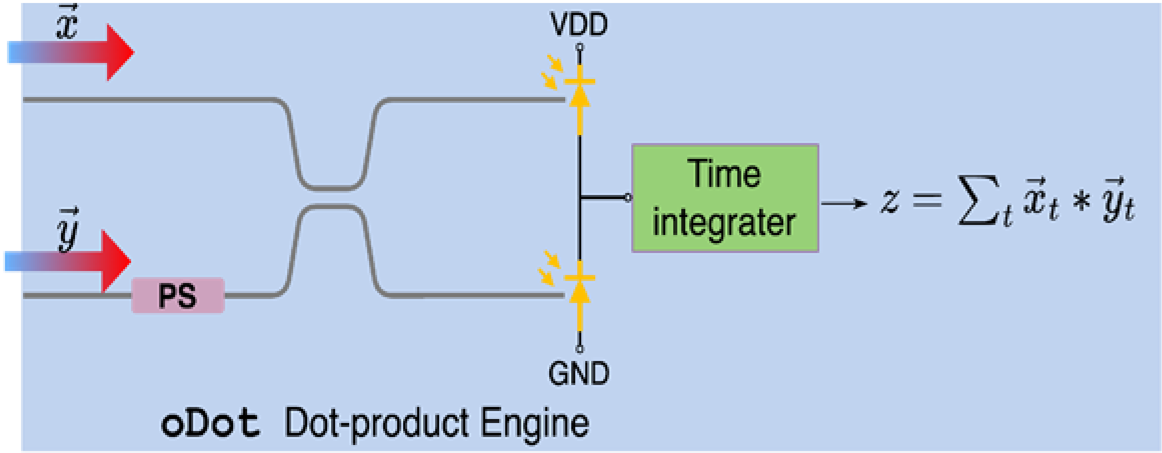
Dynamic photonic tensor core: DDTC

# Unique Arch-level Opt. in Optical Accelerator

- Optical AI efficiency bottleneck: ***Data movement*** & ***signal conversion (ADC)***
- ***Our solution:***
  - Share signals with photonic interconnects to reduce data movement cost
  - Explore data locality with a time integrator in analog domain to reduce signal conversion cost


Shared M2 Modulation Unit #$N_c$
Shared M2 Modulation Unit #1
Shared M2 Modulation Unit #0
Micro-comb
Laser
$\lambda$
#rows / #cols
DPTC $N_c$
DPTC 1
DPTC 0
Tile $0$
. . . Tile $N_t$

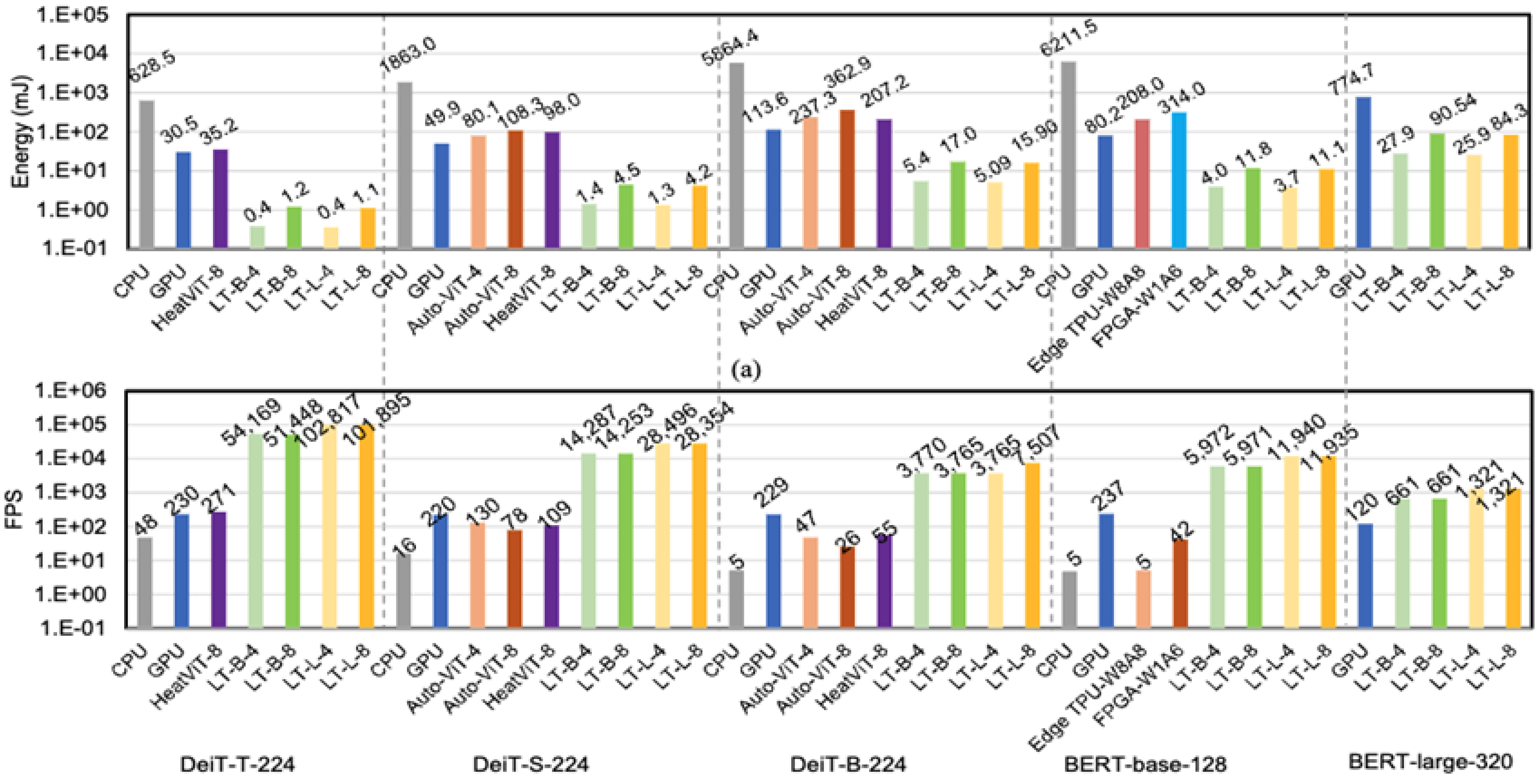
Proposed global modulation unit with optical inter-core broadcast

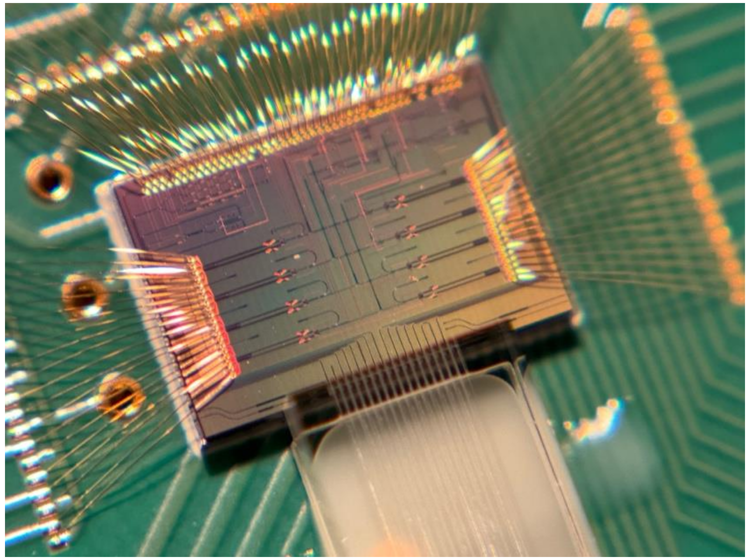


Proposed accumulation in analog domain cross time axis

# Ours vs. SOTA Digitals

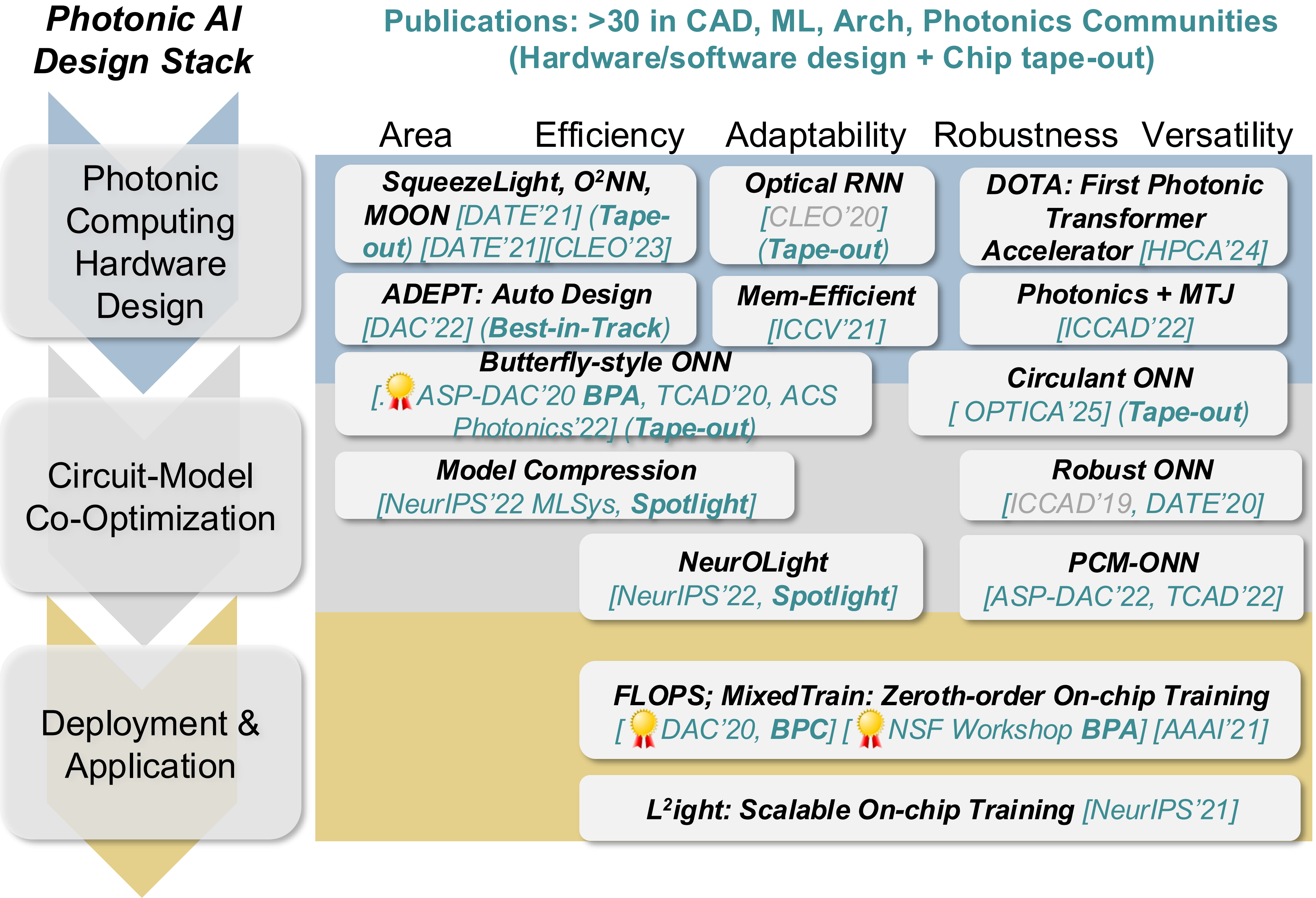
- $100 - 1000 \times$ lower energy-delay product than CPU, GPU, FPGA, Edge TPU
  - **>100x latency speedup**
- ***First*** to show the huge potential of optics for adv. ML workloads





Chip taped out
(Under testing)

# More for Our Photonics AI Journey



***Photonic AI Design Stack***

Photonic Computing Hardware Design

Circuit-Model Co-Optimization

Deployment & Application

**Publications: >30 in CAD, ML, Arch, Photonics Communities (Hardware/software design + Chip tape-out)**

Area Efficiency Adaptability Robustness Versatility

***SqueezeLight, $O^2$NN, MOON*** *[DATE'21] (**Tape-out**) [DATE'21][CLEO'23]*

***Optical RNN*** *[CLEO'20] (**Tape-out**)*

***DOTA: First Photonic Transformer Accelerator*** *[HPCA'24]*

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

***Mem-Efficient*** *[ICCV'21]*

***Photonics + MTJ*** *[ICCAD'22]*

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

***Circulant ONN*** *[ OPTICA'25] (**Tape-out**)*

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

***Robust ONN*** *[ICCAD'19, DATE'20]*

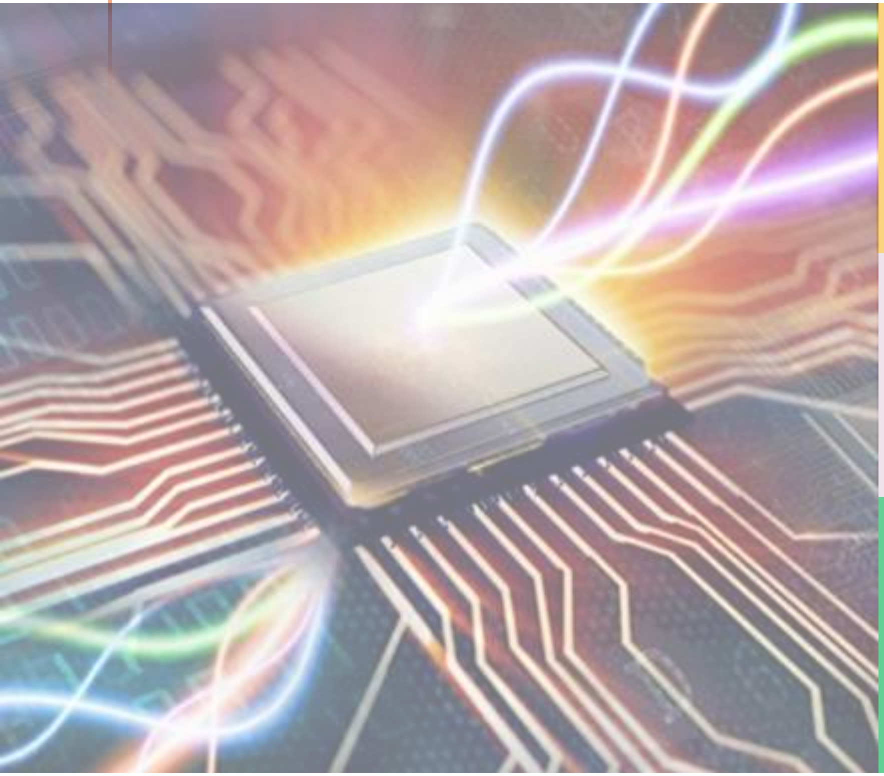
***NeurOLight*** *[NeurIPS'22, **Spotlight**]*

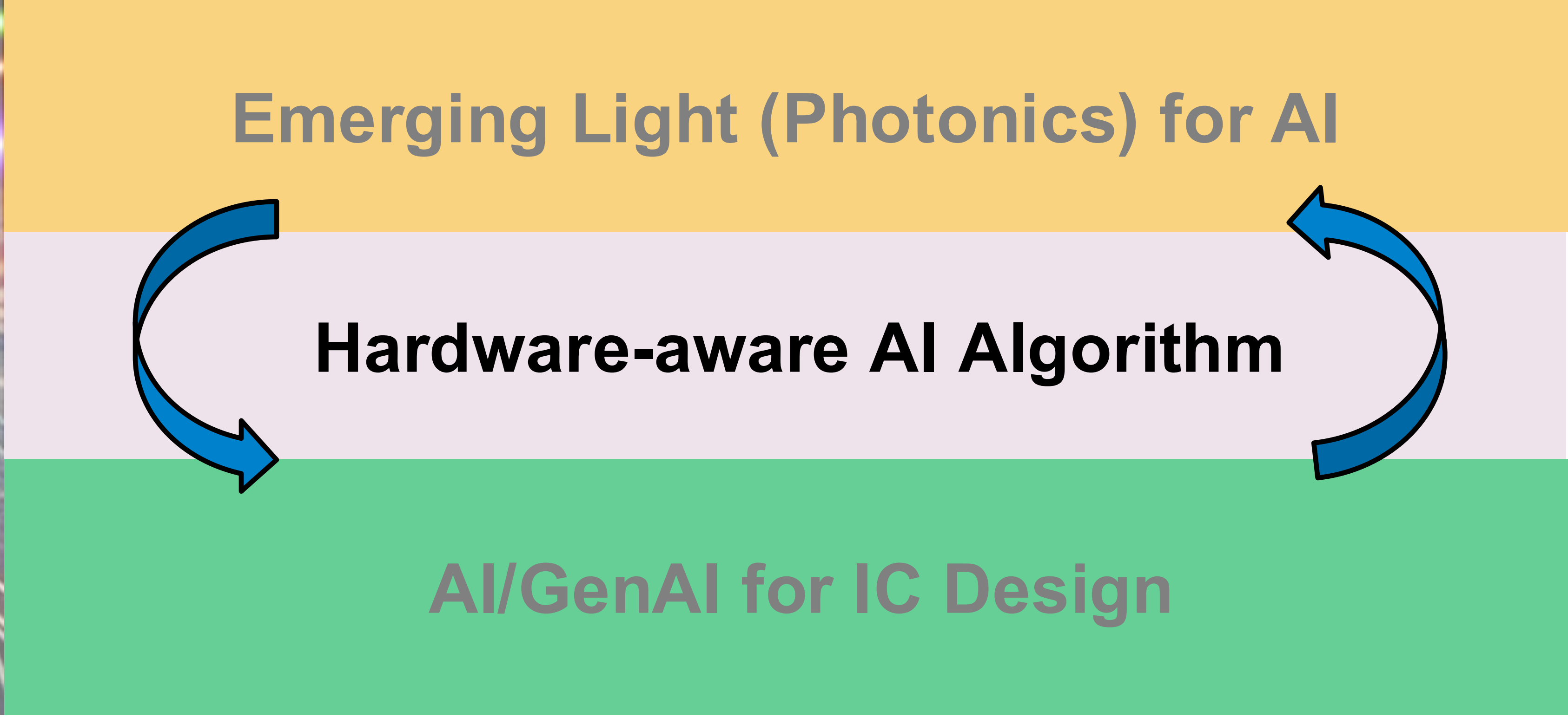
***PCM-ONN*** *[ASP-DAC'22, TCAD'22]*

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

***$L^2$ight: Scalable On-chip Training*** *[NeurIPS'21]*

# In This Talk: AI-Hardware Synergy





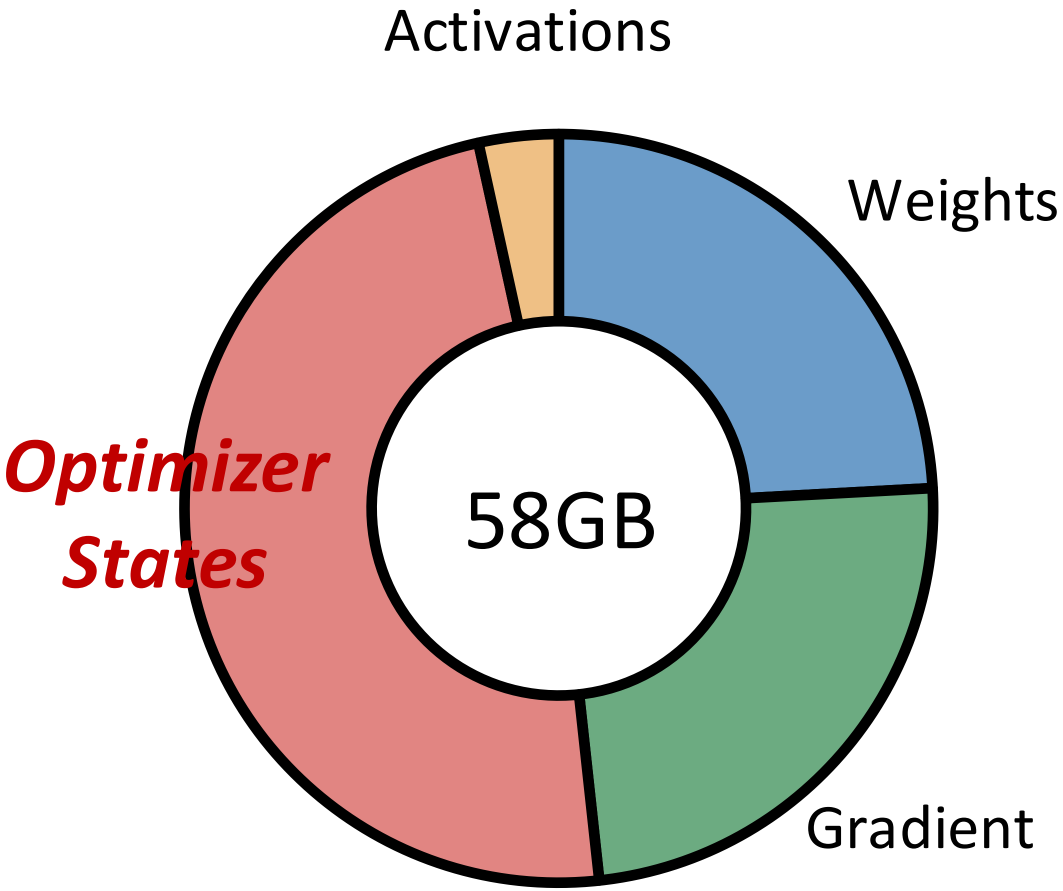
Emerging Light (Photonics) for AI

Hardware-aware AI Algorithm

AI/GenAI for IC Design

# Training of LLMs Takes a Lot of Memory!

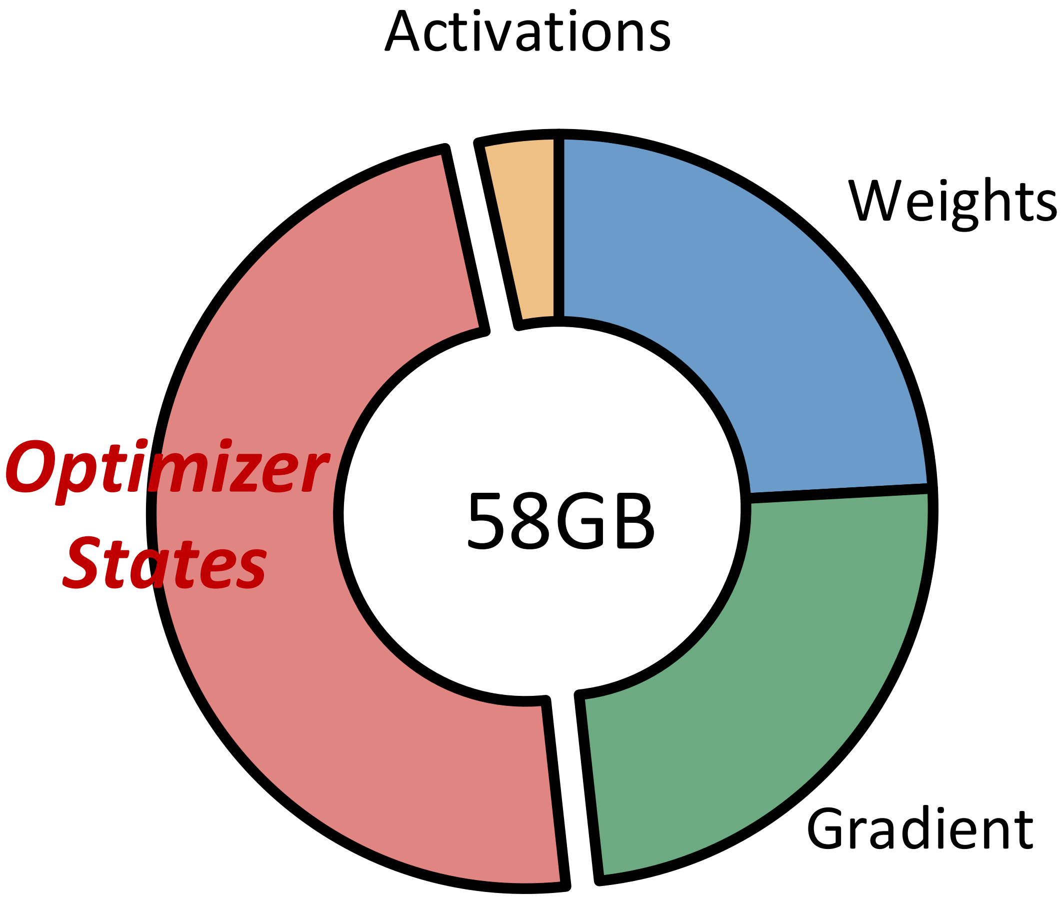
- Pre-training LLaMA-7B model (BF16, batch size of 1)
  - Trainable Parameters (Weights): 14GB
  - Gradients: 14GB
  - Activations: 2GB



# Training of LLMs Takes a Lot of Memory!

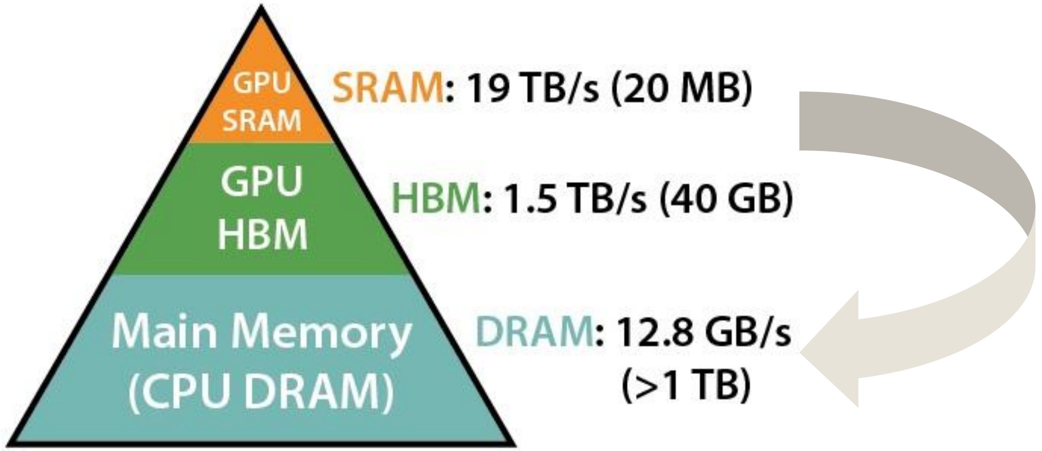
- Pre-training LLaMA-7B model (BF16, batch size of 1)
  - Trainable Parameters: 14GB
  - Gradients: 14GB
  - Activations: 2GB

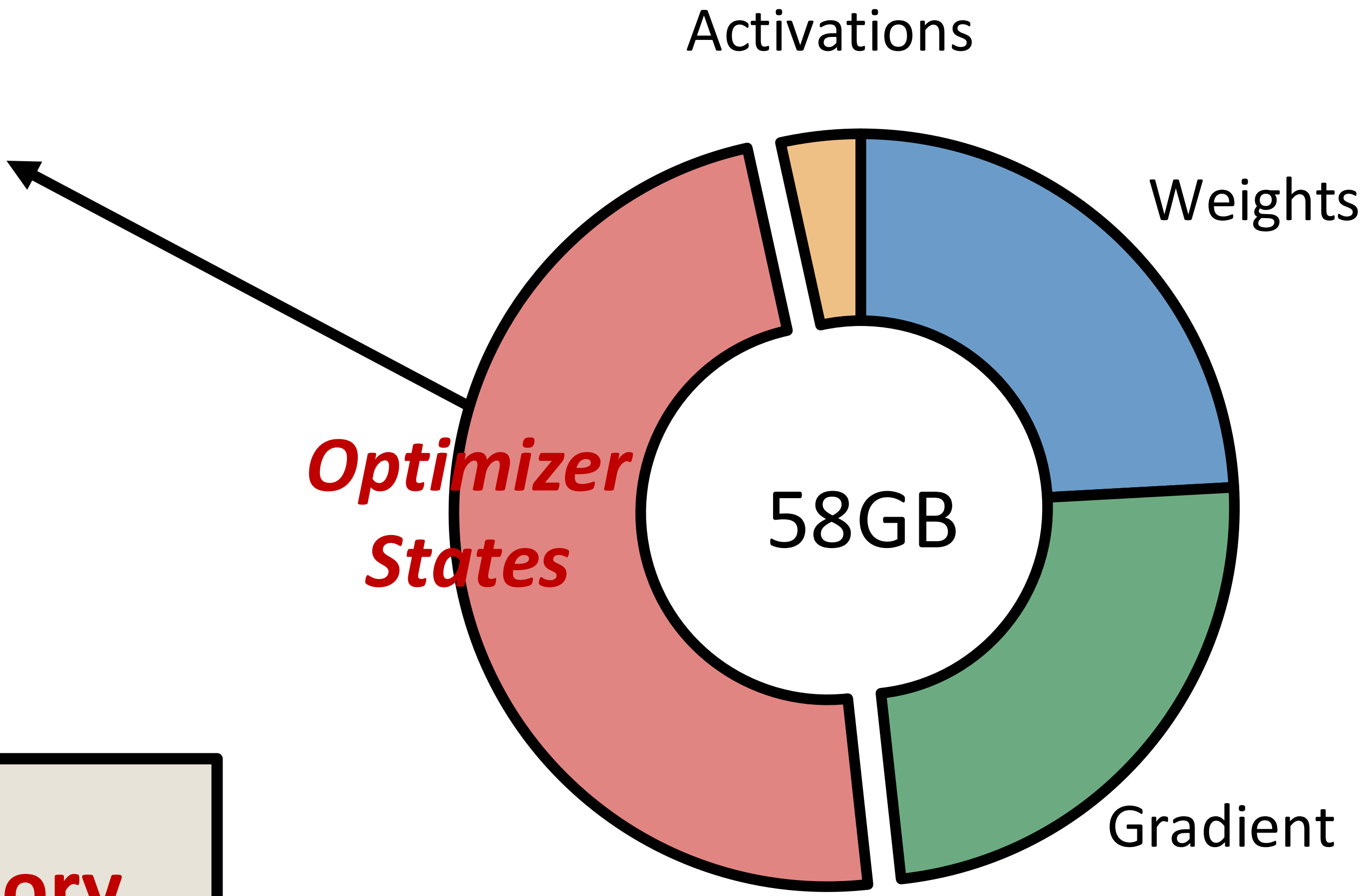
- Default optimizer→ AdamW
  - Store first and second order estimates
  - Twice of the model weights: 28GB



# Existing Solution – System-level

**Exemplar techniques:**
Memory offloading, optimizer states parallelism

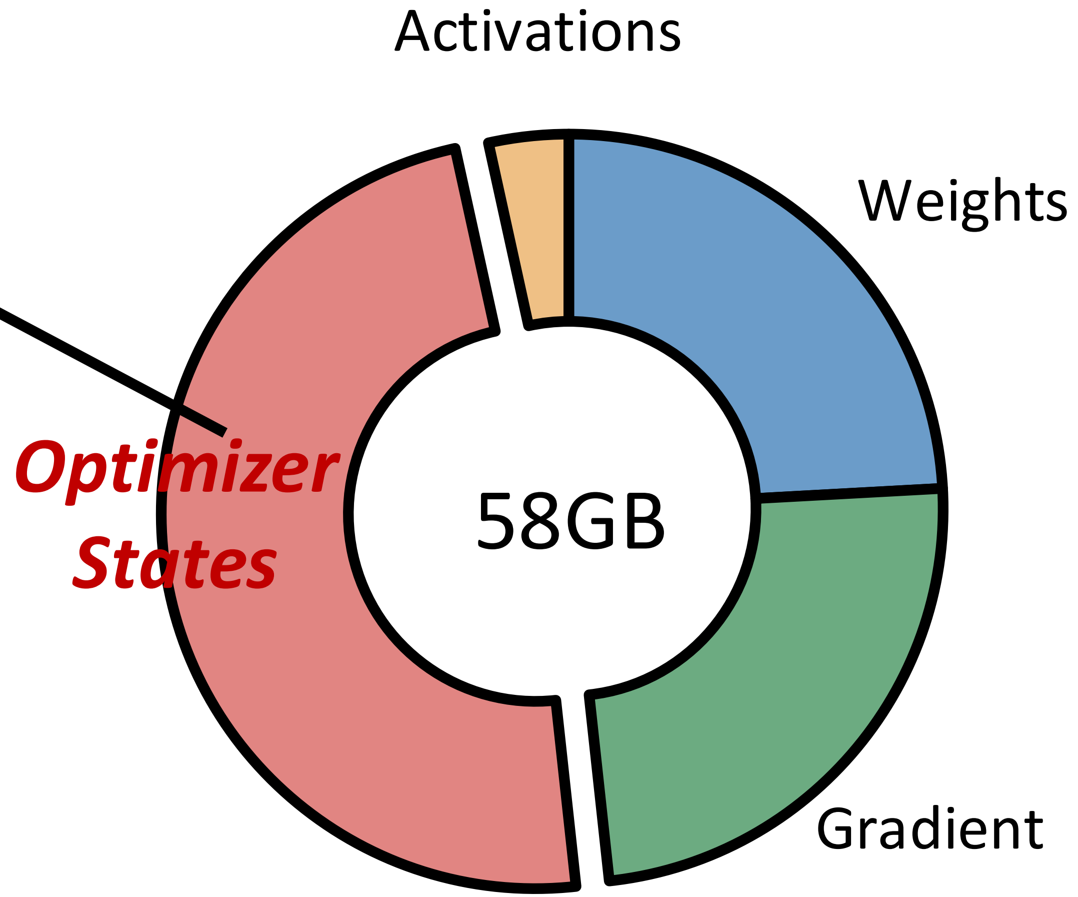




⍰ **Trade time for less memory**

# Existing Solution – Algorithm-level

**Exemplar techniques:**
Low-rank:
GaLore, Low-rank Adaption(LoRA)
Quantization:
8-bit optimizer, Low-precision Training
Optimizer Redundancy:
Adam-mini



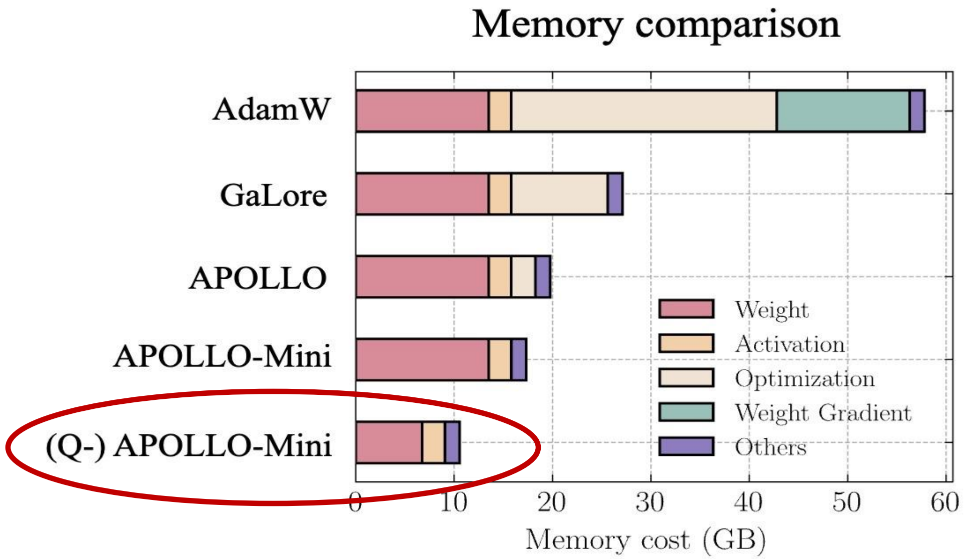
⍰ **Fine-tuning only; Still memory-intensive; Costly SVD**

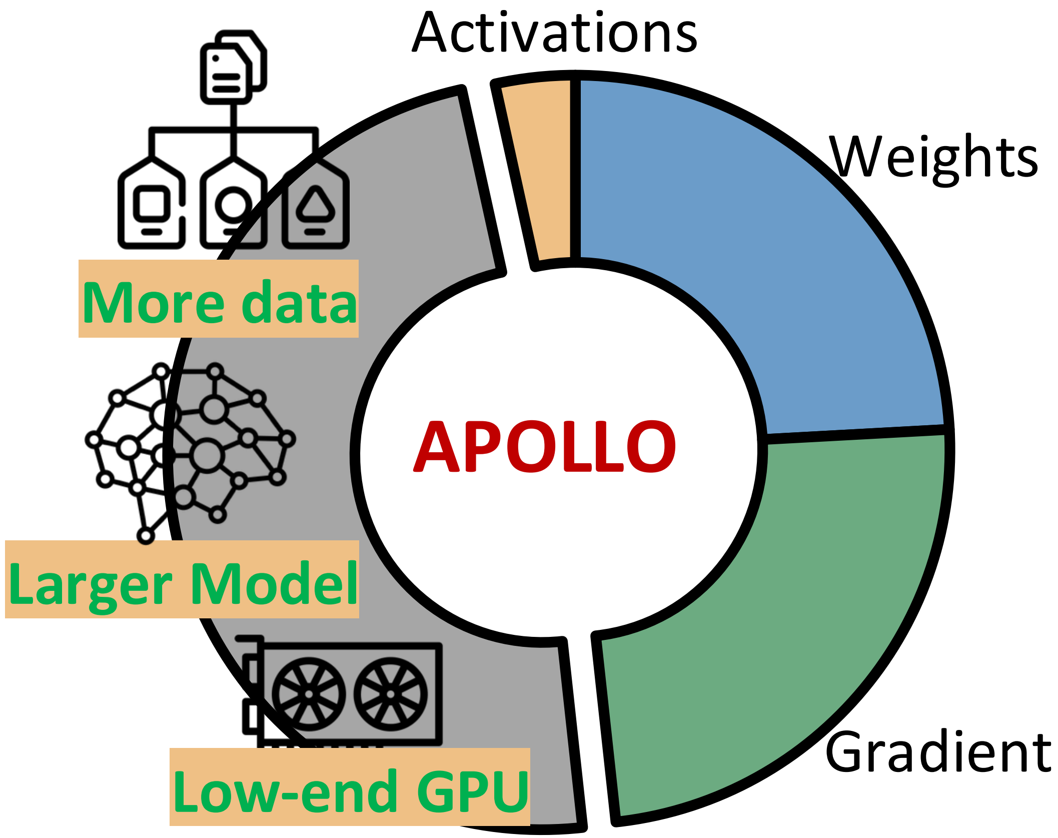
# Our Proposed APOLLO, MLSys'25

Hugging Face LLaMA-Factory Easy and Efficient LLM Fine-Tuning

flux Axolotl

- ***New-record memory-efficiency***, ***low-overhead***, ***powerful as Adam(W)***
  - Train LLaMA-7B model with <12GB memory (cf. 58GB)!!! **You can use a NVidia Titan to train 7B!**





Zhu, H., Zhang, Z., Cong, W., Liu, X., Park, S., Chandra, V., Long, B., Pan, D.Z., Wang, Z. and Lee, J., 2024. Apollo: Sgd-like memory, adamw-level performance. MLSys'25

**Outstanding Paper Honorable Mention Award**

# Rethink AdamW in a Structured Version

Adam(W)

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \cdot \tilde{\mathbf{G}}_t, \quad \tilde{\mathbf{G}}_t = \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t} + \epsilon}$$

Theoretical equivalent reformulation

Reformulated Adam(W)

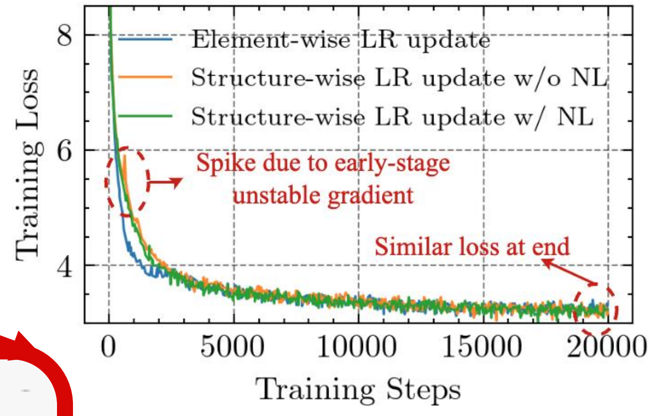
$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \cdot \boxed{\frac{\tilde{\mathbf{G}}_t}{\mathbf{G}_t}} \cdot \mathbf{G}_t$$

Structuralize $S = \frac{\tilde{G}}{G}$ into a channel-wise/tensor-wise

Structured Adam(W)

$$\tilde{\mathbf{G}}_t = \mathbf{S} \cdot \mathbf{G}_t = \mathbf{G}_t \cdot \mathrm{diag}(s).$$

$$s_j = \frac{\|\tilde{\mathbf{G}}_t[:, j]\|_2}{\|\mathbf{G}_t[:, j]\|_2}$$



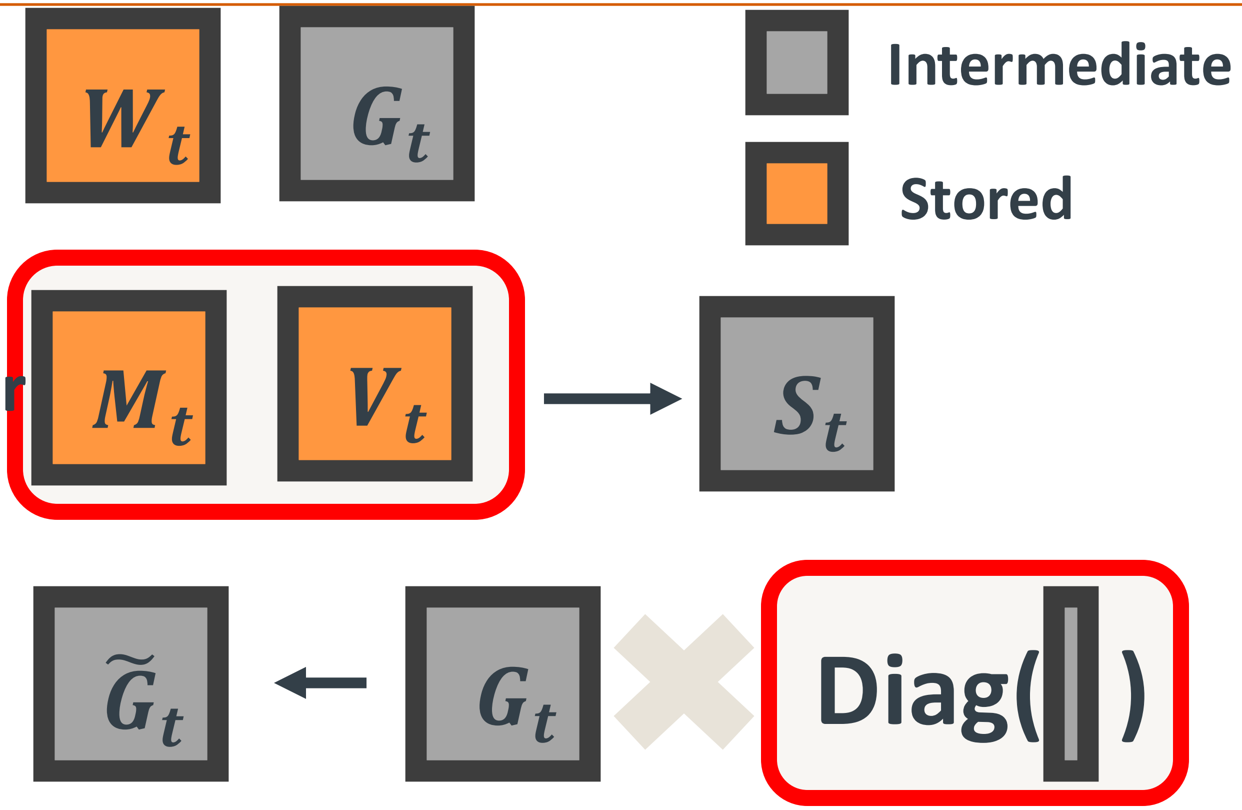
*Empirical validation on Training Loss*

# Wait 🤔! No memory benefits for doing so!

For any weight update $W_t$ at iteration t

Adam(W) updates 1st order and 2nd order moments

Obtain gradient update $\tilde{G}_t$



Structured Learning Rate Update

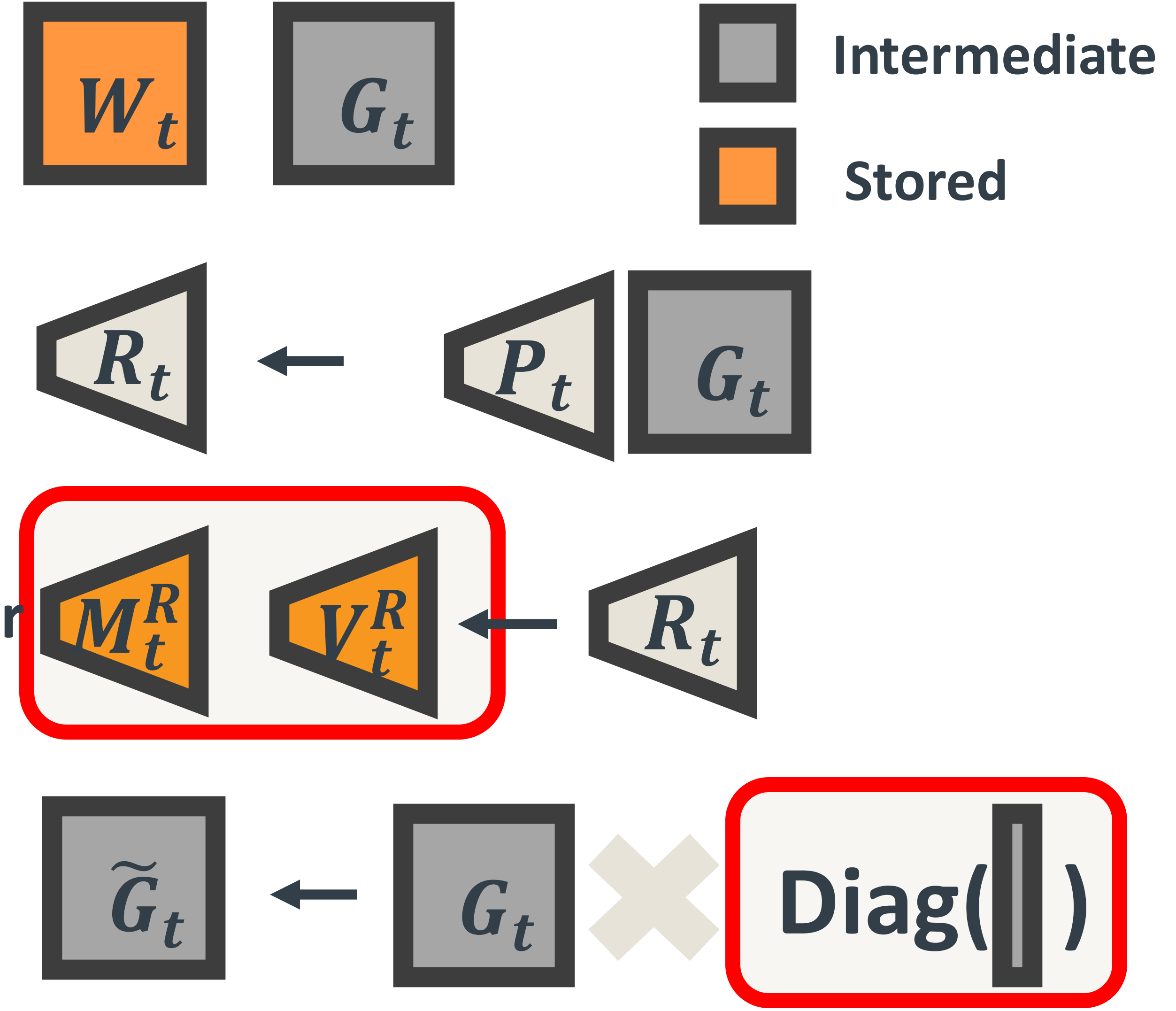
# Approximate Structured Learning Rate in Low Rank Space

For any weight update $W_t$ at iteration t

**Get the compressed gradient matrix $R_t$**

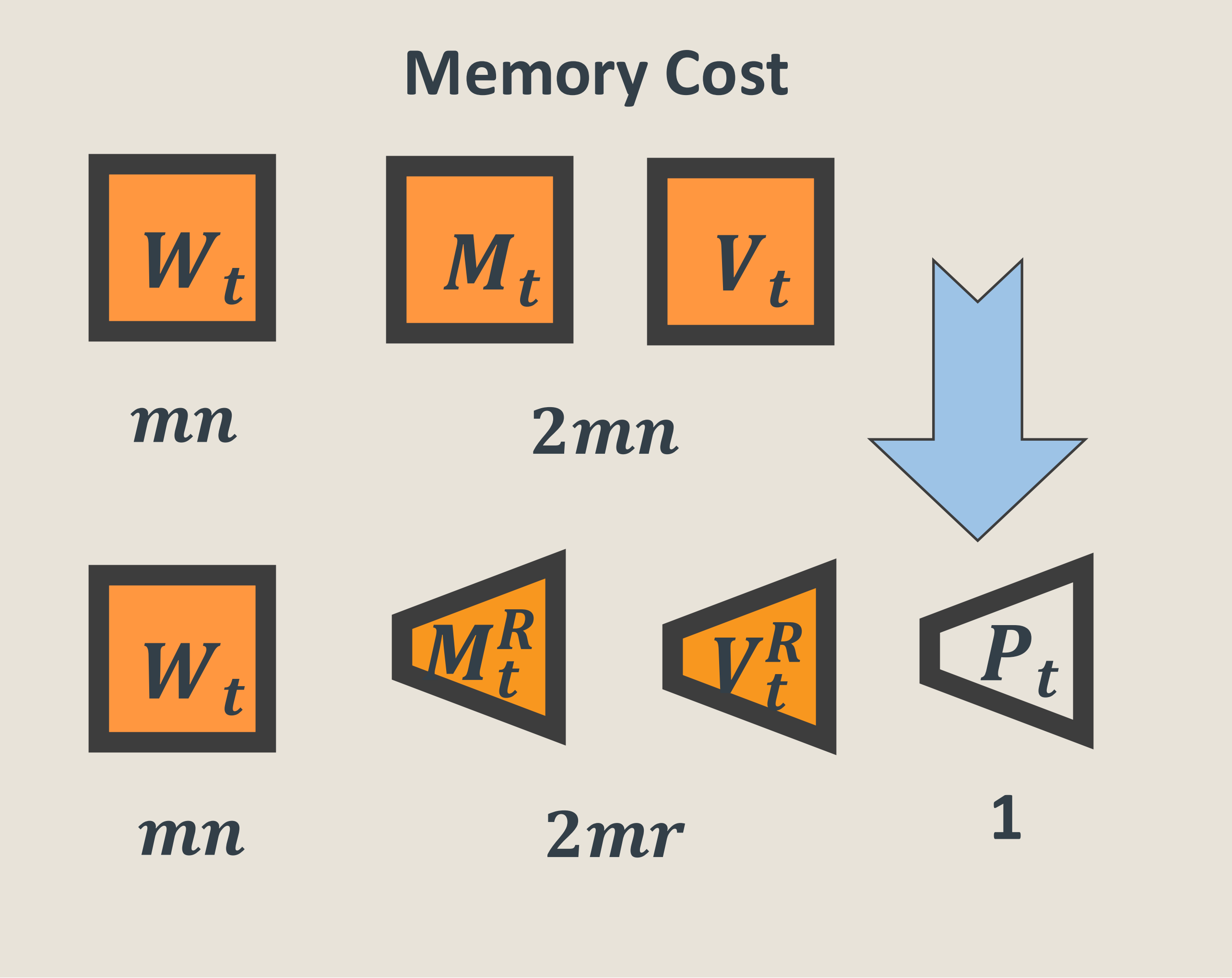
Adam(W) updates 1st order and 2nd order moments

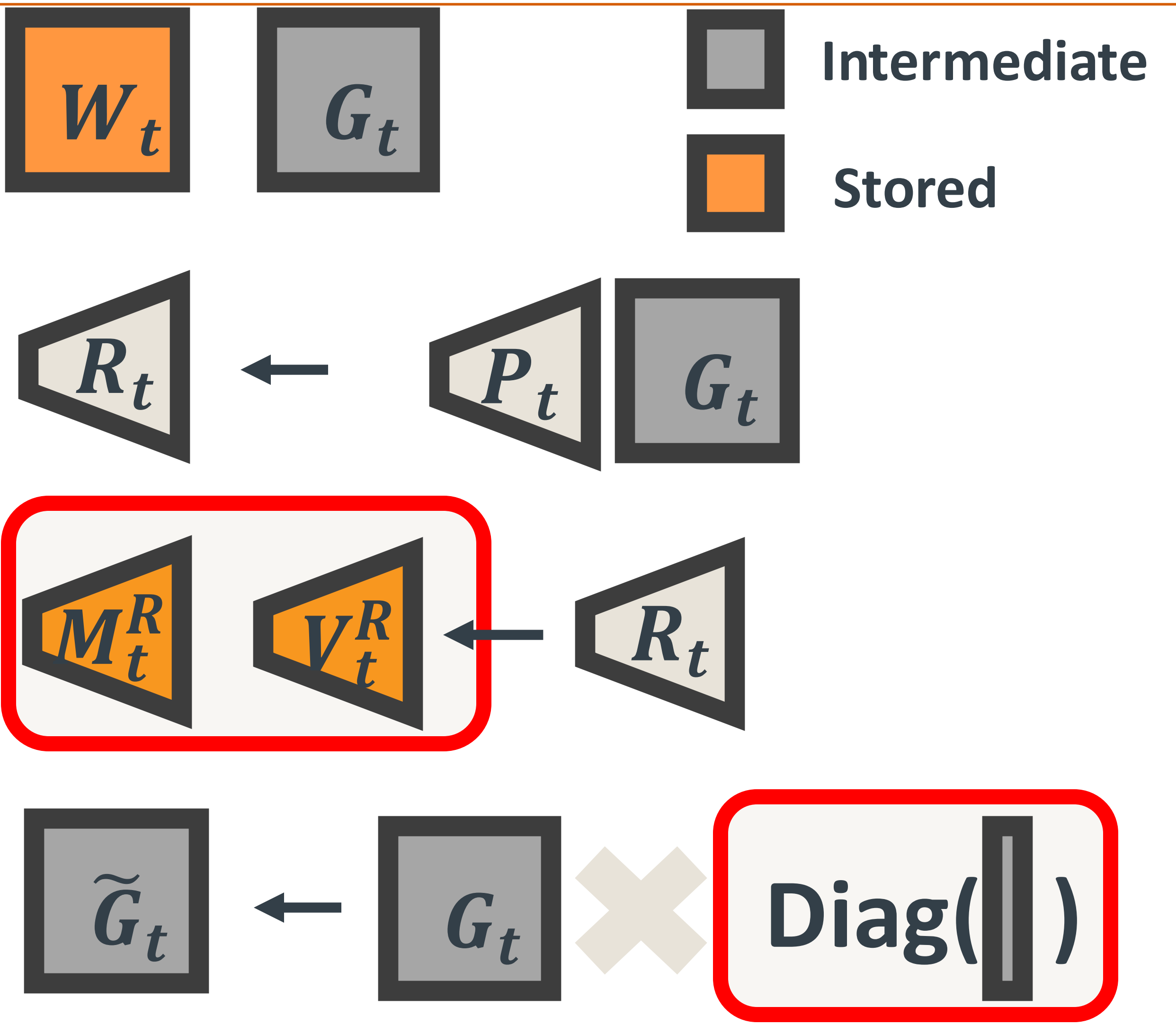
Obtain gradient update $\widetilde{G}_t$



Structured Learning Rate Update

# APOLLO: Memory benefit $2mn \rightarrow 2mr + 1$

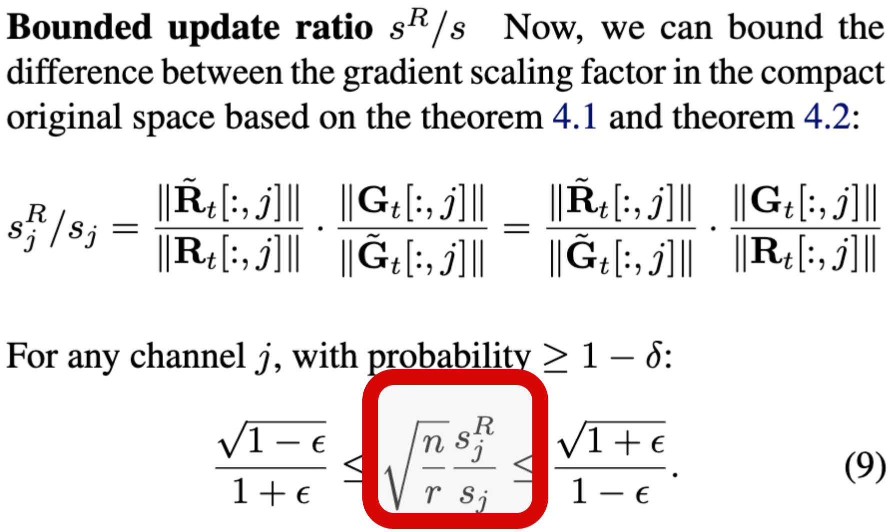



$W_t$
$G_t$
Intermediate
Stored
$R_t \leftarrow P_t \; G_t$
$M_t^R \; V_t^R \leftarrow R_t$
$\tilde{G}_t \leftarrow G_t \times$ Diag( )

Structured Learning Rate Update

# APOLLO: Many Firsts in Efficient LLM Training

- First time enable pre-training with an SVD-free approach
  - Random projection works with a theoretical bound!
  - An elegant factor to compensate error introduced by the low rank $r$

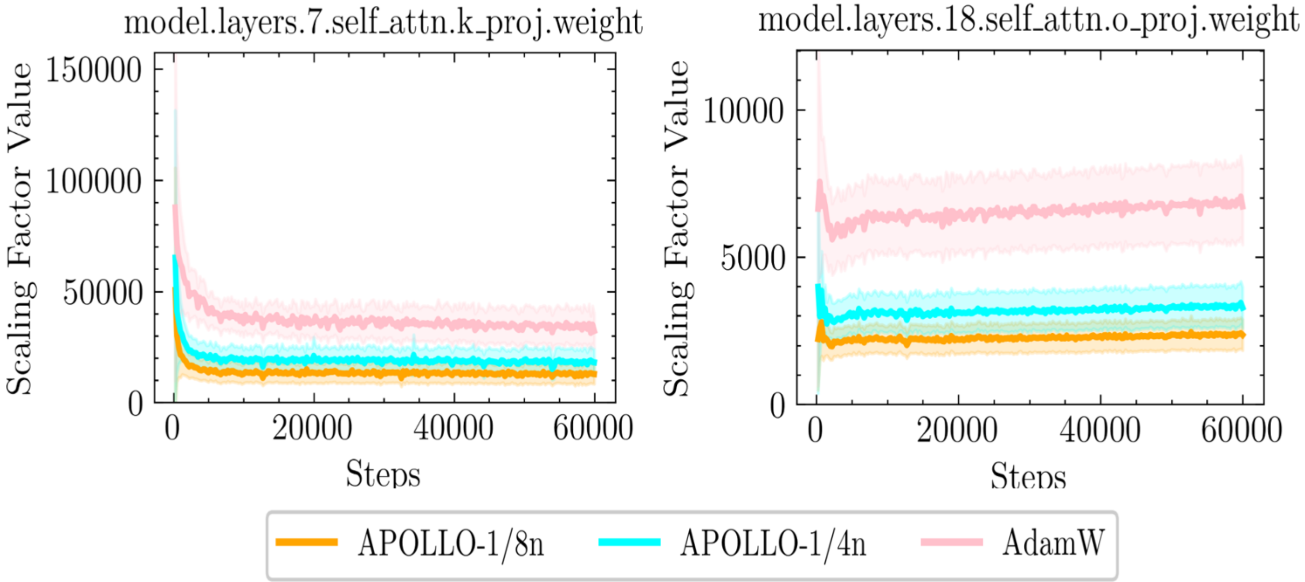


**Bounded update ratio $s^R/s$** Now, we can bound the difference between the gradient scaling factor in the compact original space based on the theorem 4.1 and theorem 4.2:

$$s_j^R/s_j = \frac{\|\tilde{\mathbf{R}}_t[:,j]\|}{\|\mathbf{R}_t[:,j]\|} \cdot \frac{\|\mathbf{G}_t[:,j]\|}{\|\tilde{\mathbf{G}}_t[:,j]\|} = \frac{\|\tilde{\mathbf{R}}_t[:,j]\|}{\|\tilde{\mathbf{G}}_t[:,j]\|} \cdot \frac{\|\mathbf{G}_t[:,j]\|}{\|\mathbf{R}_t[:,j]\|}$$

For any channel $j$, with probability $\geq 1-\delta$:

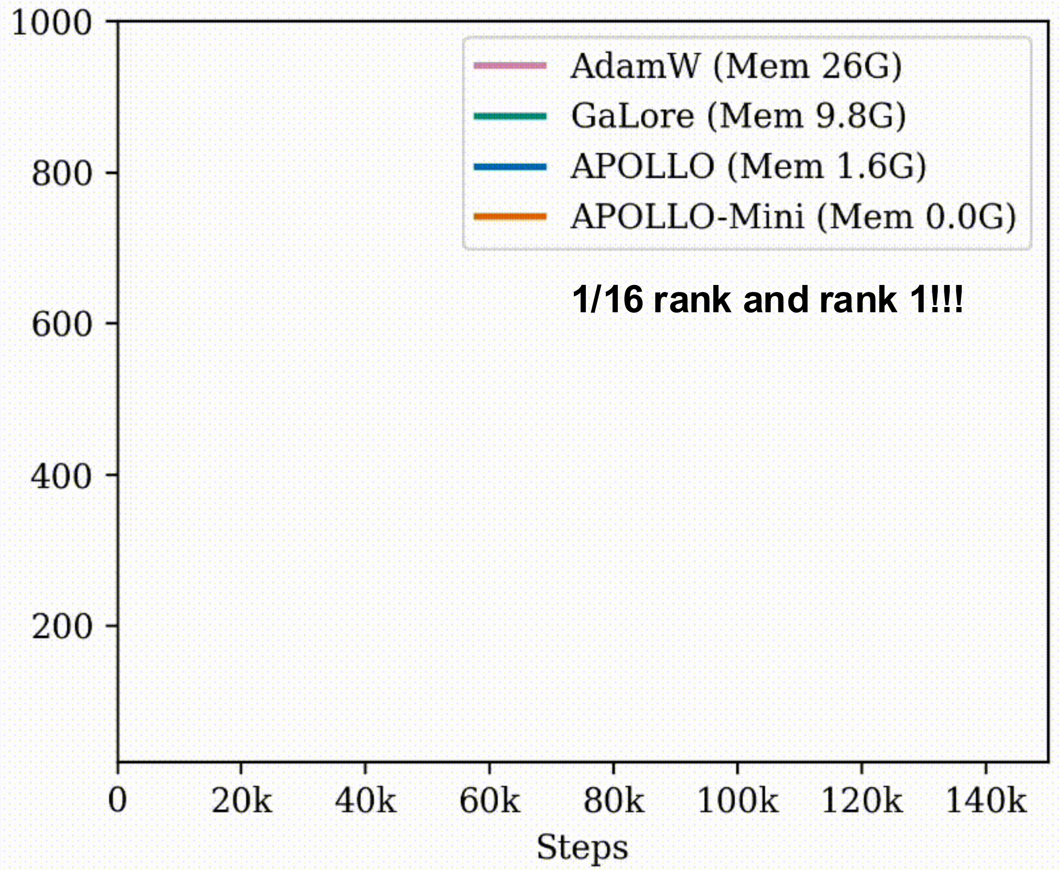
$$\frac{\sqrt{1-\epsilon}}{1+\epsilon} \leq \boxed{\sqrt{\frac{n}{r}\frac{s_j^R}{s_j}}} \leq \frac{\sqrt{1+\epsilon}}{1-\epsilon}. \tag{9}$$



- First time enable pre-training with only rank 1 space (r=1), using tensor-wise scaling

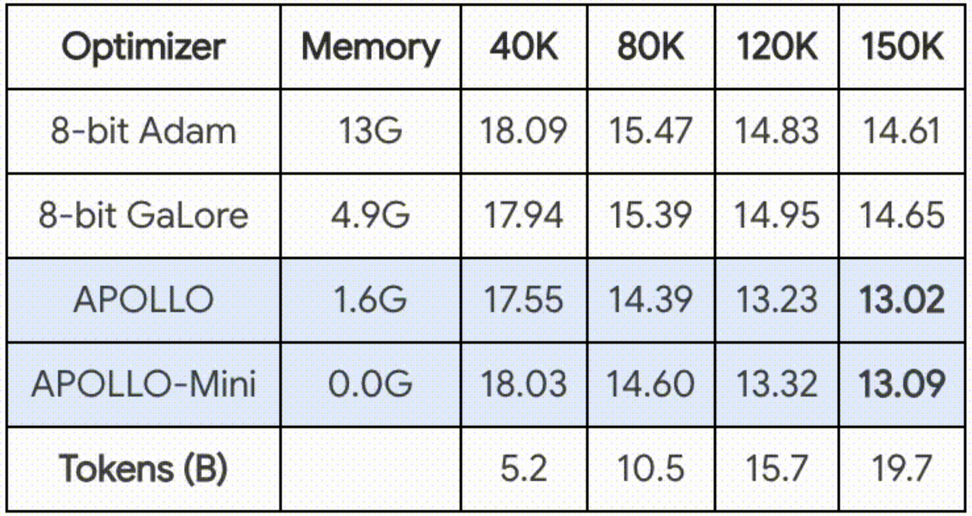
# Performance & Throughput: Pre-training LLaMA 7B

- **On-par or even better** than AdamW even at 1/16 rank and rank 1!!!
- First to finish 7B training in 2 weeks (**3x faster than Adam**)



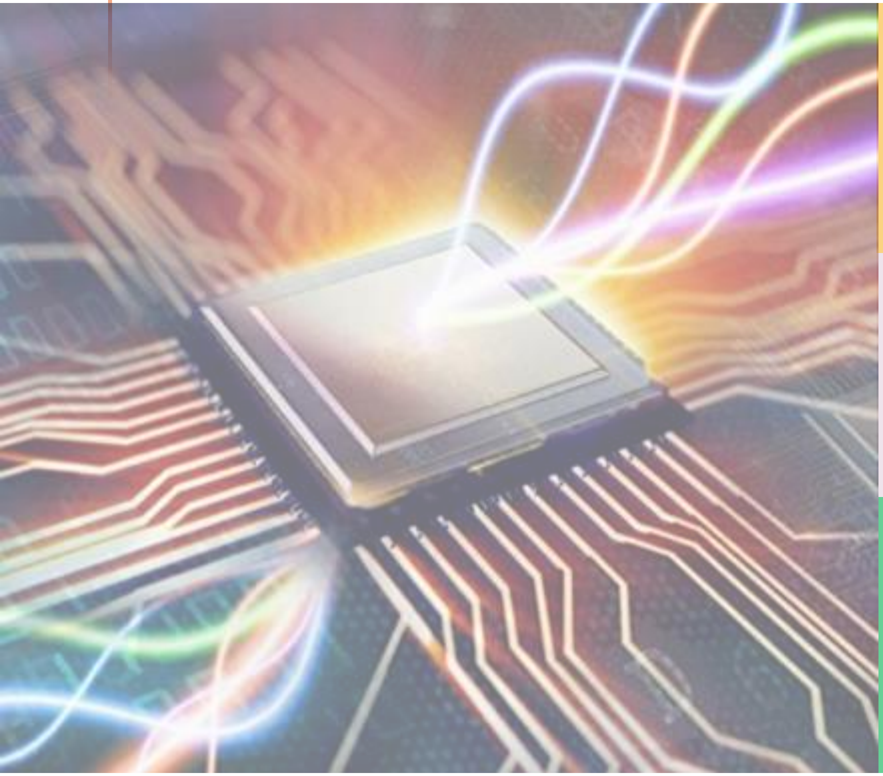
**1/16 rank and rank 1!!!**

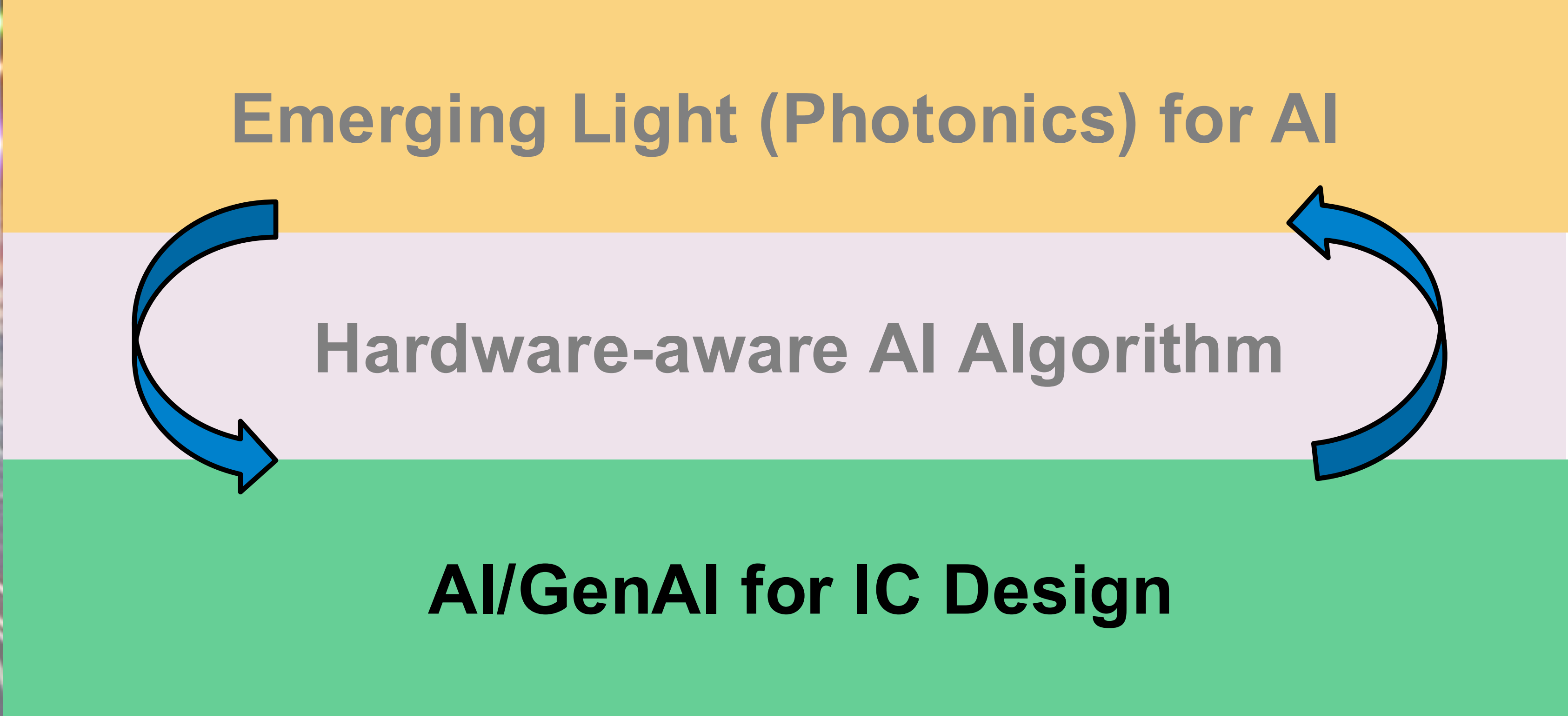
Pre-training LLaMA 7B on C4 dataset for 150K steps with reported perplexity



| Optimizer | Memory | 40K | 80K | 120K | 150K |
|---|---|---|---|---|---|
| 8-bit Adam | 13G | 18.09 | 15.47 | 14.83 | 14.61 |
| 8-bit GaLore | 4.9G | 17.94 | 15.39 | 14.95 | 14.65 |
| APOLLO | 1.6G | 17.55 | 14.39 | 13.23 | **13.02** |
| APOLLO-Mini | 0.0G | 18.03 | 14.60 | 13.32 | **13.09** |
| **Tokens (B)** | | 5.2 | 10.5 | 15.7 | 19.7 |

# In This Talk: AI-Hardware Synergy





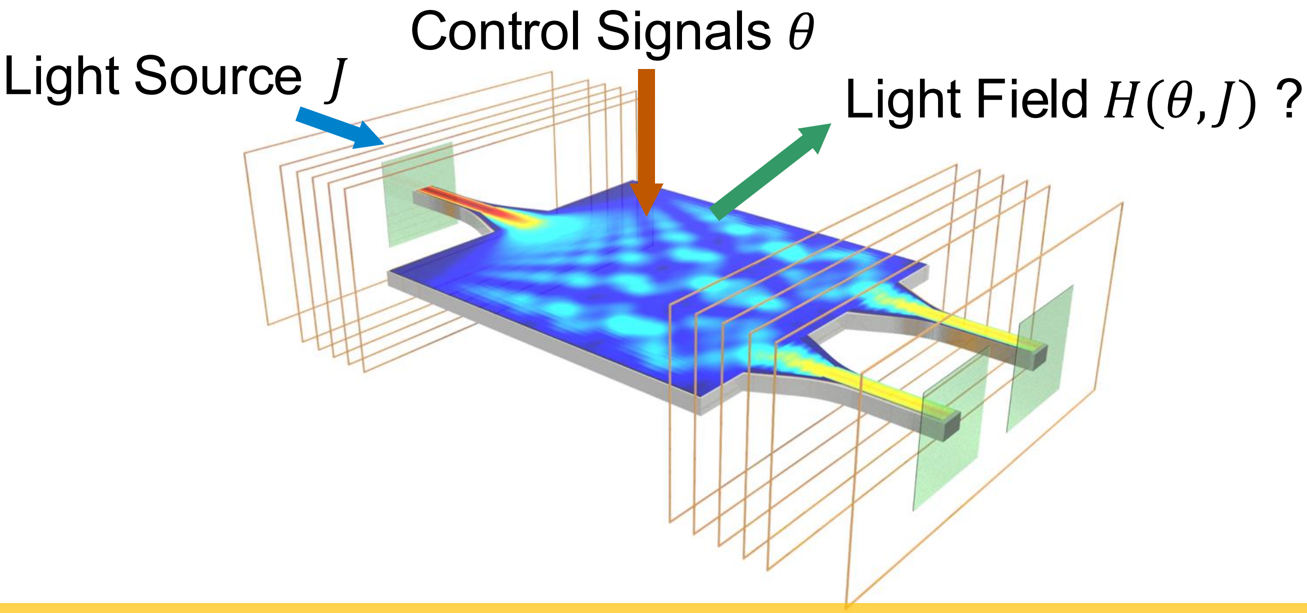
Emerging Light (Photonics) for AI

Hardware-aware AI Algorithm

**AI/GenAI for IC Design**

# AI-Assisted Simulations for Optical Designs

- Optical AI has great potential with customized structures ➔ novel optical devices
- However, computationally expensive simulations for Maxwell equations, etc.

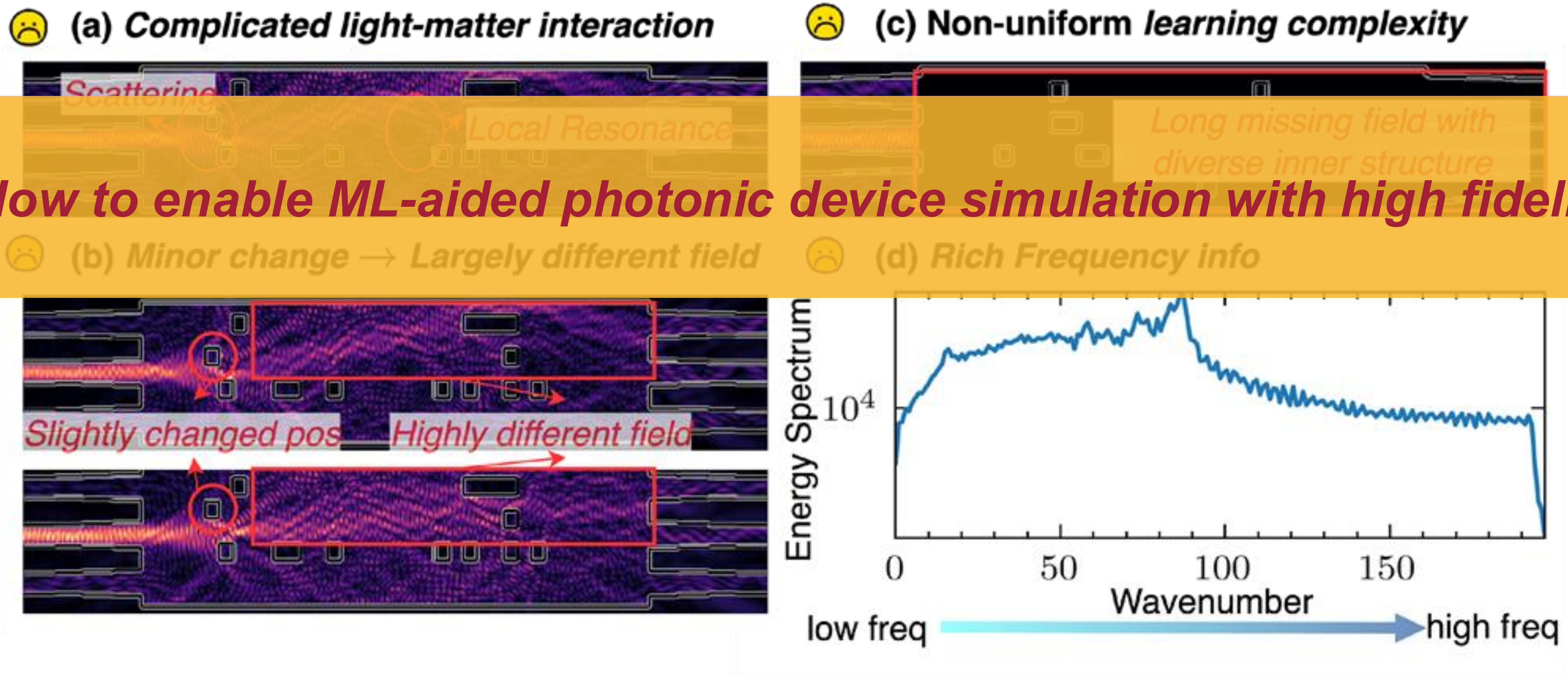


**Can ML models learn the light propagation principles?**

**→ Fast AI-based Maxwell Solver → novel optical device design**

# Complicated PDE for Real-world Device

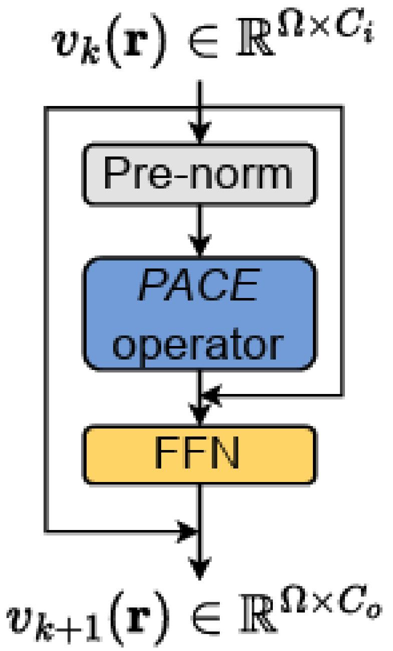
- ML for PDE has been popular to speed up simulation process
- But not an easy task for real-world photonic devices



***How to enable ML-aided photonic device simulation with high fidelity?***

# Our Proposed PACE: New Operator Kernel

- **A math-inspired neural operator kernel**
  - Better computation and parameter efficiency than Attention as the kernel

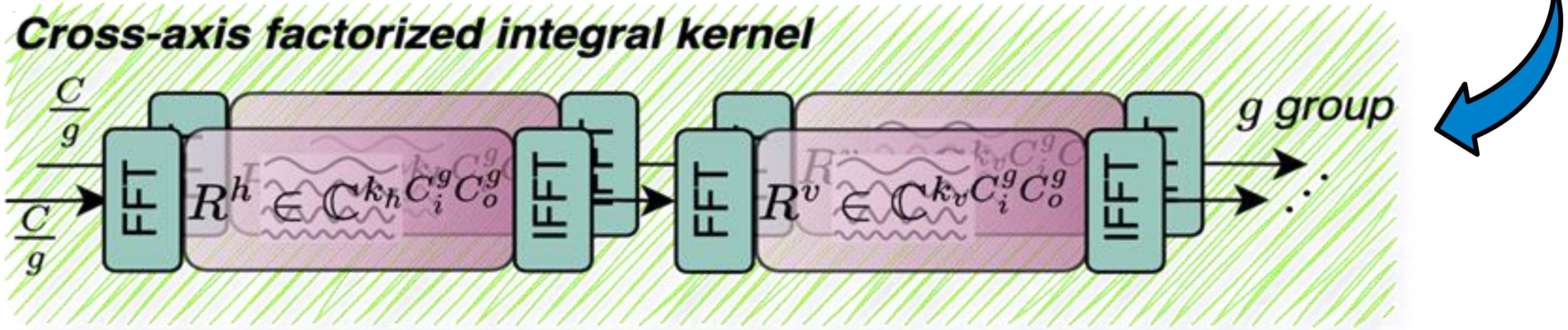


$v_k(\mathbf{r}) \in \mathbb{R}^{\Omega\times C_i}$
Pre-norm
PACE
operator
FFN
$v_{k+1}(\mathbf{r}) \in \mathbb{R}^{\Omega\times C_o}$

*Transformer-style Design*

$$\mathcal{K}v_k)(\boldsymbol{r}_1) = \int_{\Omega} \kappa(\boldsymbol{r}_1,\boldsymbol{r}_2)v_k(\boldsymbol{r}_2)\mathrm{d}v_k(\boldsymbol{r}_2), \quad \forall \boldsymbol{r}_1 \in \Omega,$$

$$\approx \int_{\Omega_h} \kappa(\boldsymbol{r}_1,\boldsymbol{r}_2)^h \int_{\Omega_h} \kappa(\boldsymbol{r}_1,\boldsymbol{r}_2)^v v_k(\boldsymbol{r}_2)\mathrm{d}v_k(\boldsymbol{r}_2)^v \mathrm{d}v_k(\boldsymbol{r}_2)^h, \quad \forall \boldsymbol{r}_1 \in \Omega$$



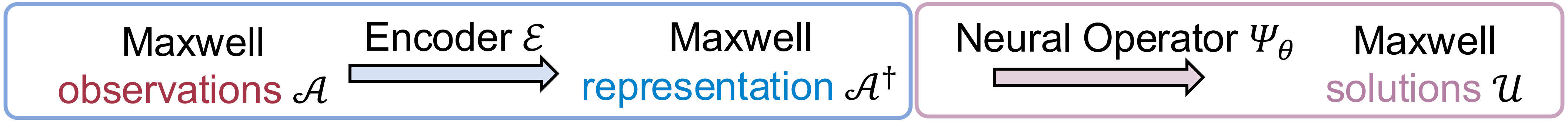
Cross-axis factorized integral kernel
$\frac{C}{g}$
$\frac{C}{g}$
FFT
$R^h \in \mathbb{C}^{k_h C_i^g C_o^g}$
IFFT
FFT
$R^v \in \mathbb{C}^{k_v C_i^g C_o^g}$
IFFT
g group

*A math-inspired neural operator kernel for approximated 2D integral*

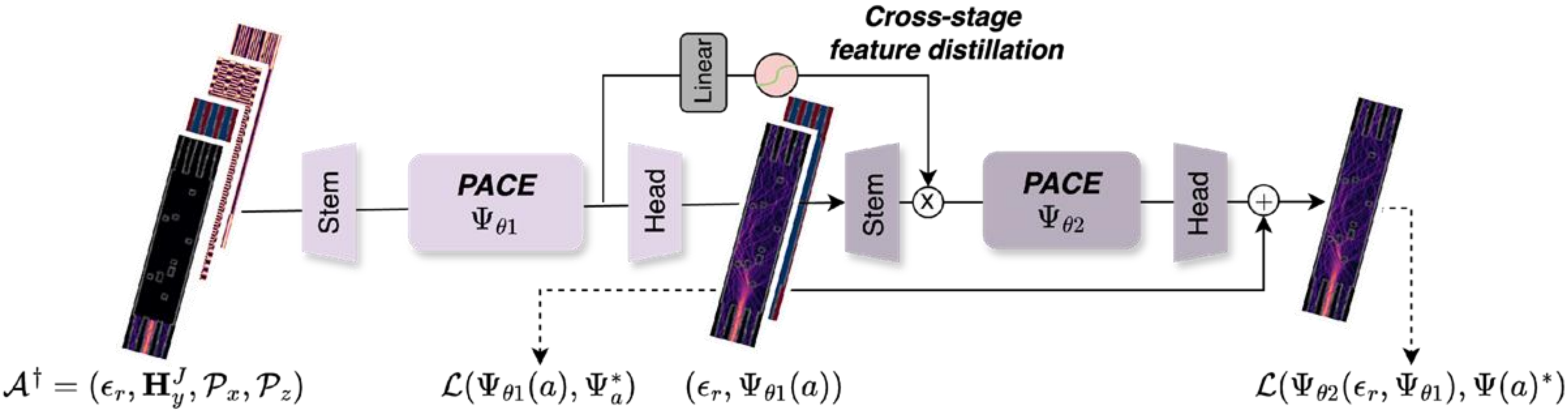
Zhu, Hanqing, Wenyan Cong, Guojin Chen, Shupeng Ning, Ray Chen, Jiaqi Gu, and David Z. Pan. "Pace: Pacing operator learning to accurate optical field simulation for complicated photonic devices.", NeurIPS 2024

# Our Proposed PACE: Learning from Rough to Clear

- ♦ ***A math-inspired neural operator kernel + New training recipe***



Maxwell observations $\mathcal{A}$
Encoder $\mathcal{E}$
Maxwell representation $\mathcal{A}^\dagger$
Neural Operator $\Psi_\theta$
Maxwell solutions $\mathcal{U}$



Cross-stage feature distillation
Linear
Stem
PACE $\Psi_{\theta 1}$
Head
Stem
PACE $\Psi_{\theta 2}$
Head
$\mathcal{A}^\dagger = (\epsilon_r, \mathbf{H}_y^J, \mathcal{P}_x, \mathcal{P}_z)$
$\mathcal{L}(\Psi_{\theta 1}(a), \Psi_a^*)$
$(\epsilon_r, \Psi_{\theta 1}(a))$
$\mathcal{L}(\Psi_{\theta 2}(\epsilon_r, \Psi_{\theta 1}), \Psi(a)^*)$

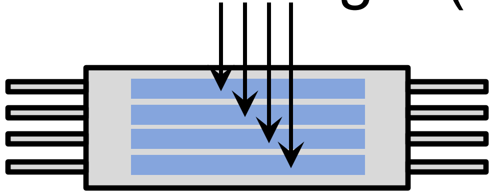
*Stage I: Learn a rough solution only from Maxwell observations*

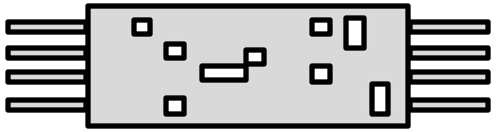
*Stage II: Learn a fine solution from the rough optical field solution*

Zhu, Hanqing, Wenyan Cong, Guojin Chen, Shupeng Ning, Ray Chen, Jiaqi Gu, and David Z. Pan. "Pace: Pacing operator learning to accurate optical field simulation for complicated photonic devices.", NeurIPS 2024

# Main Results

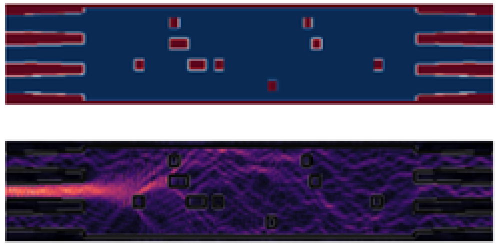
- ♦ Benchmarks: subwavelength (etched) MMIs and Metaline





- ♦ PACE: A much stronger baseline for photonic simulation

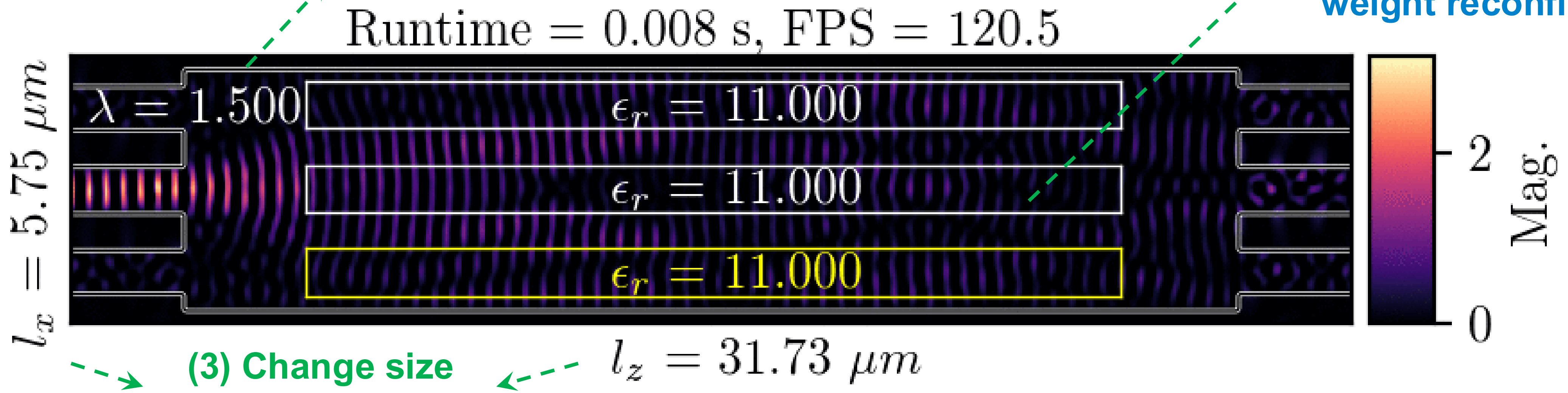
**53.8%** lower error and **50%** fewer parameters

| Benchmarks | Model | #Params (M) ↓ | Train Err ($10^{-2}$) ↓ | Test Err ($10^{-2}$) ↓ |
|---|---|---|---|---|
| Etched MMI 3x3<br> | UNet [20, 4] | 3.88 | 63.03 | 65.32 |
| | Dil-ResNet [28] | 4.17 | 51.34 | 51.79 |
| | Attention-based model [18] | 3.75 | 70.05 | 69.85 |
| | U-NO [2] | 4.38 | 34.22 | 42.86 |
| | Latent-spectral method [36] | 4.81 | 55.07 | 55.16 |
| | FNO-2d [19] | 3.99 | 32.51 | 38.71 |
| | Tensorized FNO-2d [16] | 2.25 | 35.52 | 36.61 |
| | Factorized FNO-2d [32] | 4.02 | 24.2 | 32.81 |
| | NeurOLight [10] | 2.11 | 15.58 | 17.21 |
| | PACE | **1.71** | **9.51** | **10.59** |

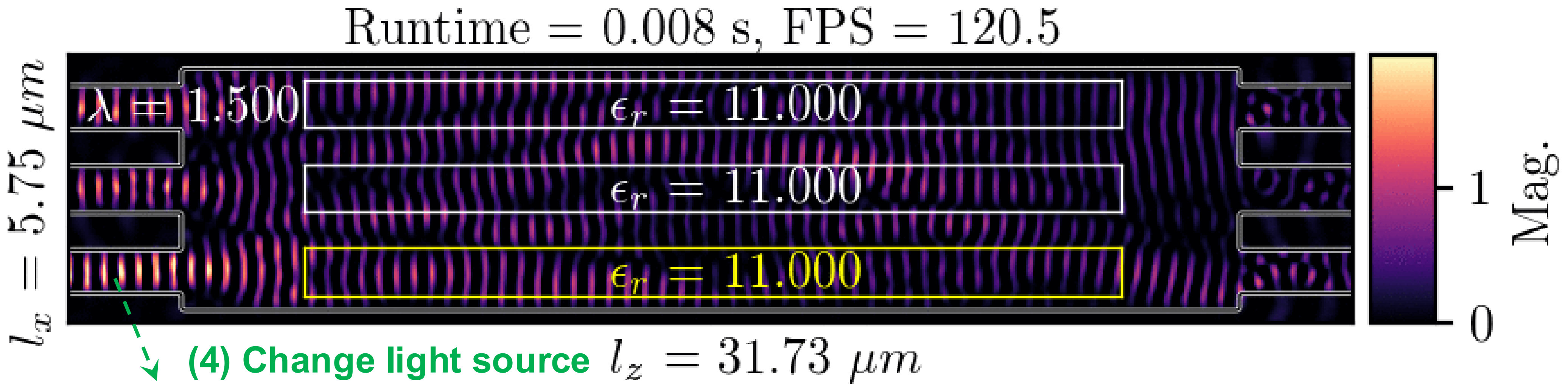
# Real-time Optical Field Prediction

(2) Change wavelength for spectrum sweeping

(1) Change index for weight reconfig.

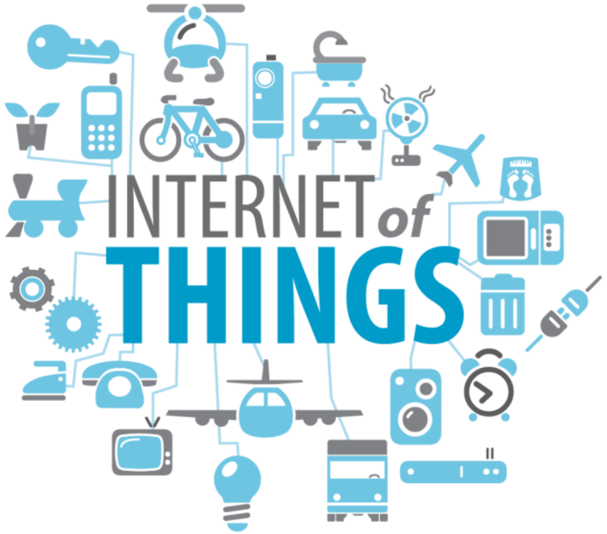


(3) Change size for device opt.

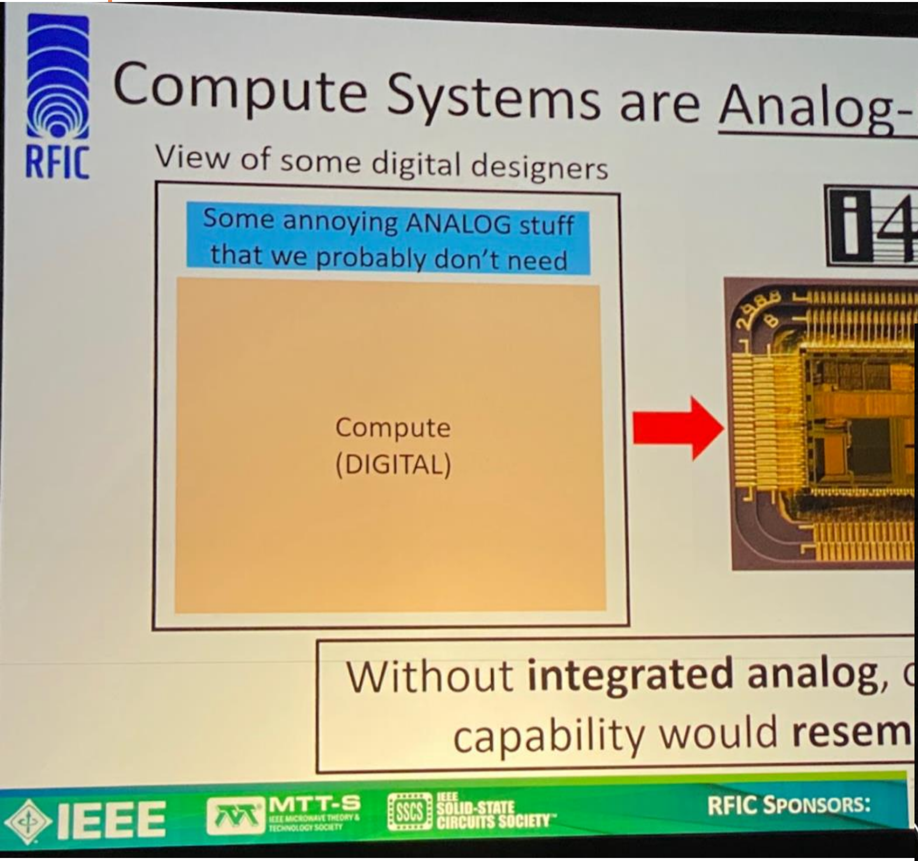


(4) Change light source as different inputs

# Growing Analog/RF IC Demand



# Even Digitals are Analog-Enabled!



## Compute Systems are Analog-Enabled Digital

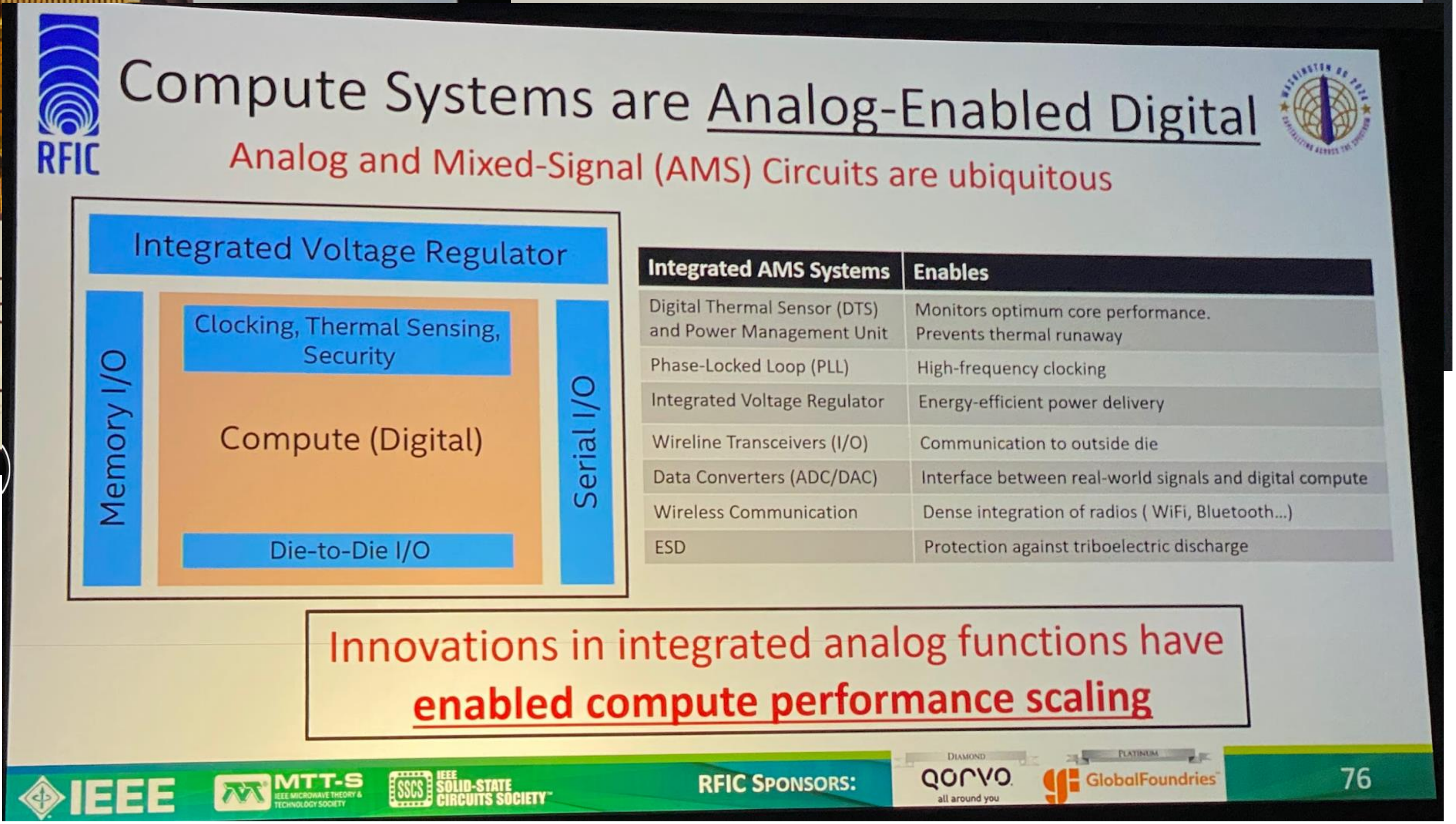
View of some digital designers

Some annoying ANALOG stuff that we probably don't need

Compute (DIGITAL)

Without **integrated analog**, c
capability would **resem**

Analog Evolution in Digital CMOS-Centric Technologies



## Compute Systems are Analog-Enabled Digital

Analog and Mixed-Signal (AMS) Circuits are ubiquitous

Integrated Voltage Regulator

Memory I/O

Clocking, Thermal Sensing, Security

Compute (Digital)

Die-to-Die I/O

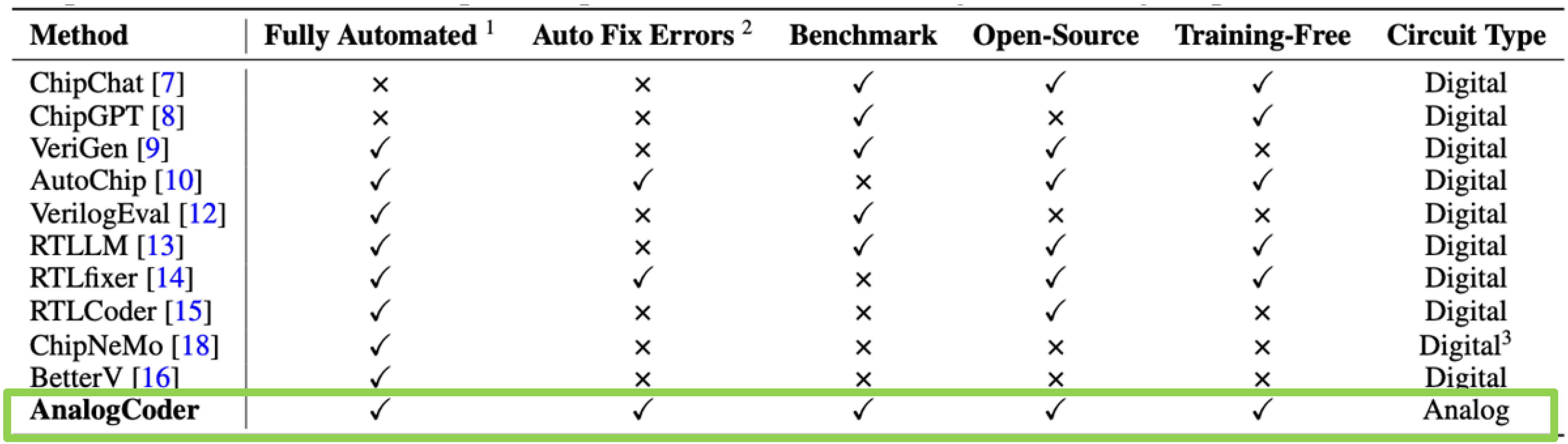
Serial I/O

| Integrated AMS Systems | Enables |
|---|---|
| Digital Thermal Sensor (DTS) and Power Management Unit | Monitors optimum core performance. Prevents thermal runaway |
| Phase-Locked Loop (PLL) | High-frequency clocking |
| Integrated Voltage Regulator | Energy-efficient power delivery |
| Wireline Transceivers (I/O) | Communication to outside die |
| Data Converters (ADC/DAC) | Interface between real-world signals and digital compute |
| Wireless Communication | Dense integration of radios ( WiFi, Bluetooth...) |
| ESD | Protection against triboelectric discharge |

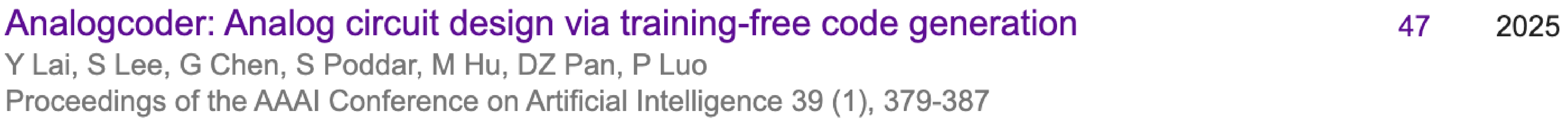
Innovations in integrated analog functions have **enabled compute performance scaling**

Prof. Tsu-Jae Liu's RFIC 2024 Keynote

# AnalogCoder: Analog Circuit Design via LLM



| Method | Fully Automated [1] | Auto Fix Errors [2] | Benchmark | Open-Source | Training-Free | Circuit Type |
|---|---|---|---|---|---|---|
| ChipChat [7] | × | × | ✓ | ✓ | ✓ | Digital |
| ChipGPT [8] | × | × | ✓ | × | ✓ | Digital |
| VeriGen [9] | ✓ | × | ✓ | ✓ | × | Digital |
| AutoChip [10] | ✓ | ✓ | × | ✓ | ✓ | Digital |
| VerilogEval [12] | ✓ | × | ✓ | × | × | Digital |
| RTLLM [13] | ✓ | × | ✓ | ✓ | ✓ | Digital |
| RTLfixer [14] | ✓ | ✓ | × | ✓ | ✓ | Digital |
| RTLCoder [15] | ✓ | × | × | ✓ | × | Digital |
| ChipNeMo [18] | ✓ | × | × | × | × | Digital[3] |
| BetterV [16] | ✓ | × | × | × | × | Digital |
| **AnalogCoder** | ✓ | ✓ | ✓ | ✓ | ✓ | Analog |



Analogcoder: Analog circuit design via training-free code generation 47 2025

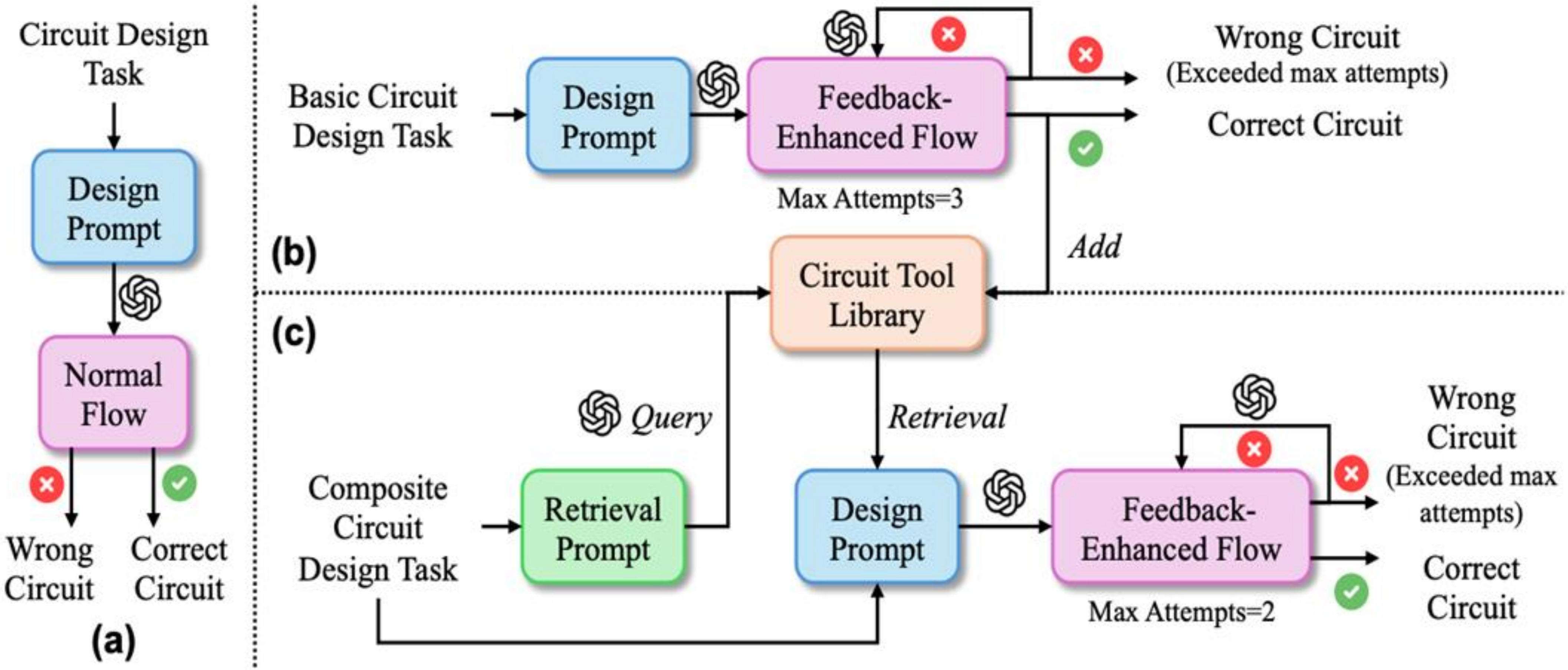
Y Lai, S Lee, G Chen, S Poddar, M Hu, DZ Pan, P Luo

Proceedings of the AAAI Conference on Artificial Intelligence 39 (1), 379-387

**AAAI 2025 Oral (< 5% acceptance rate), already got 47 citations!**

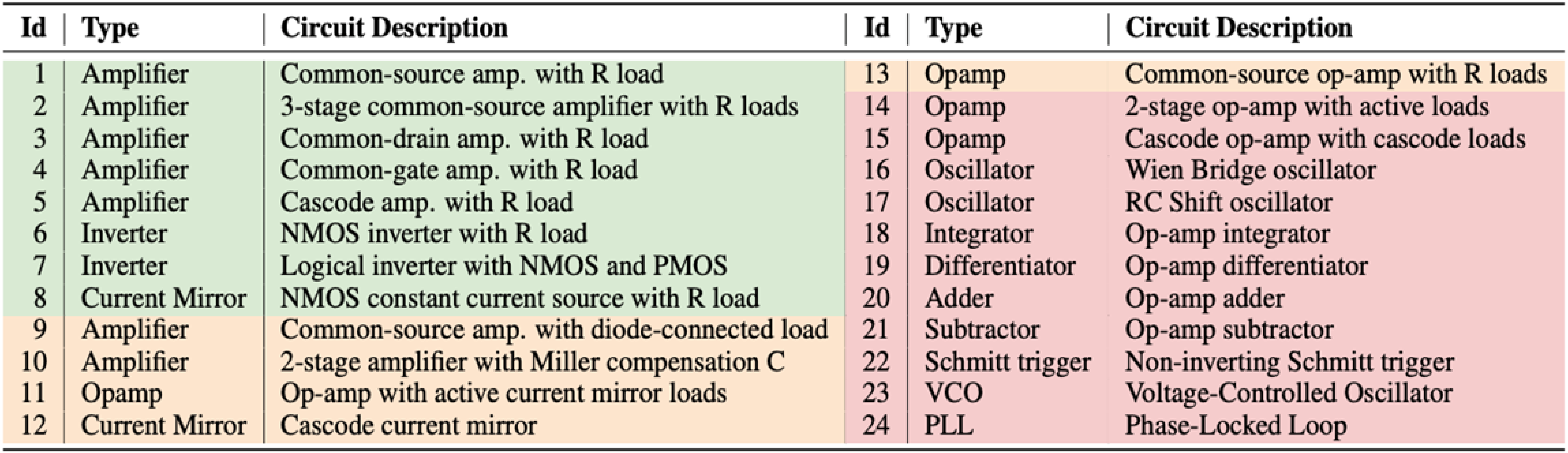
**Open sourced: https://github.com/laiyao1/AnalogCoder**

# AnalogCoder Design Flow



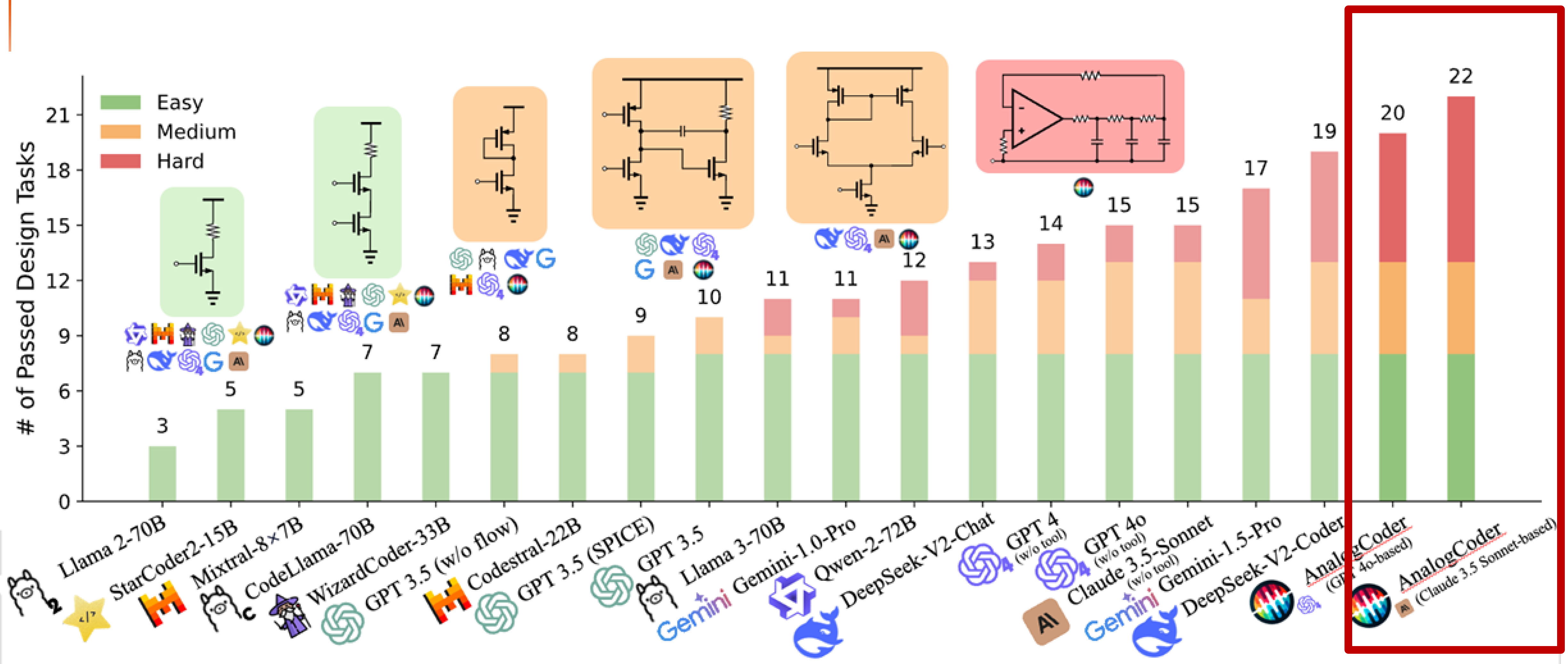
# Benchmark Circuits

- We created a set of analog circuits for benchmarking
- Amplifier, Inverter, Current Mirror, Oscillator, Integrator, …
- **Easy** / **Medium** / **Hard**



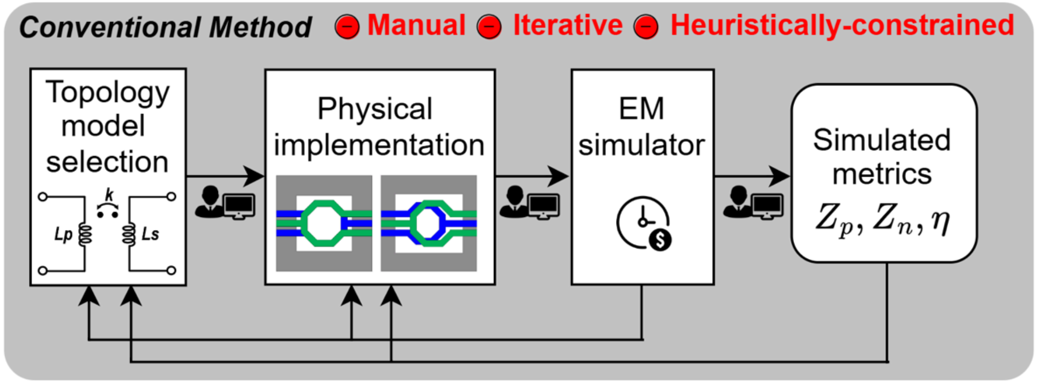
| Id | Type | Circuit Description | Id | Type | Circuit Description |
|---|---|---|---|---|---|
| 1 | Amplifier | Common-source amp. with R load | 13 | Opamp | Common-source op-amp with R loads |
| 2 | Amplifier | 3-stage common-source amplifier with R loads | 14 | Opamp | 2-stage op-amp with active loads |
| 3 | Amplifier | Common-drain amp. with R load | 15 | Opamp | Cascode op-amp with cascode loads |
| 4 | Amplifier | Common-gate amp. with R load | 16 | Oscillator | Wien Bridge oscillator |
| 5 | Amplifier | Cascode amp. with R load | 17 | Oscillator | RC Shift oscillator |
| 6 | Inverter | NMOS inverter with R load | 18 | Integrator | Op-amp integrator |
| 7 | Inverter | Logical inverter with NMOS and PMOS | 19 | Differentiator | Op-amp differentiator |
| 8 | Current Mirror | NMOS constant current source with R load | 20 | Adder | Op-amp adder |
| 9 | Amplifier | Common-source amp. with diode-connected load | 21 | Subtractor | Op-amp subtractor |
| 10 | Amplifier | 2-stage amplifier with Miller compensation C | 22 | Schmitt trigger | Non-inverting Schmitt trigger |
| 11 | Opamp | Op-amp with active current mirror loads | 23 | VCO | Voltage-Controlled Oscillator |
| 12 | Current Mirror | Cascode current mirror | 24 | PLL | Phase-Locked Loop |

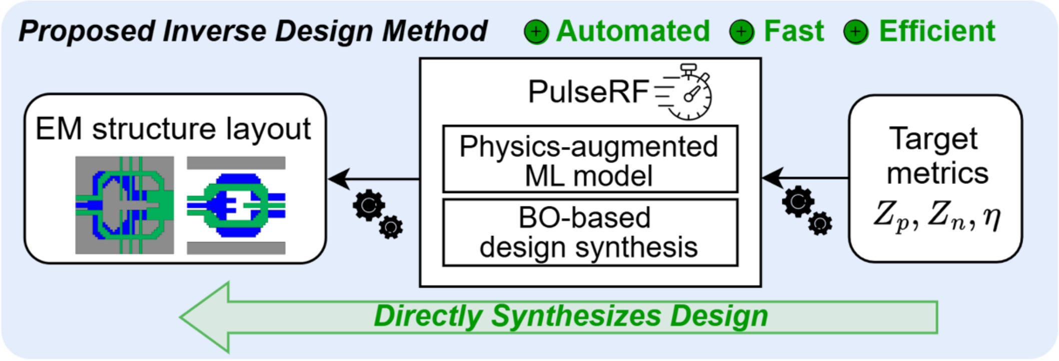
# Leaderboard of LLMs for Analog Design



# PulseRF for RFIC Passive Design

- Conventional vs. our **PulseRF** approach [Chae+, ICCAD'24]




Proposed Inverse Design Method
⊕ Automated ⊕ Fast ⊕ Efficient
EM structure layout
PulseRF
Physics-augmented ML model
BO-based design synthesis
Target metrics
$Z_p, Z_n, \eta$
Directly Synthesizes Design

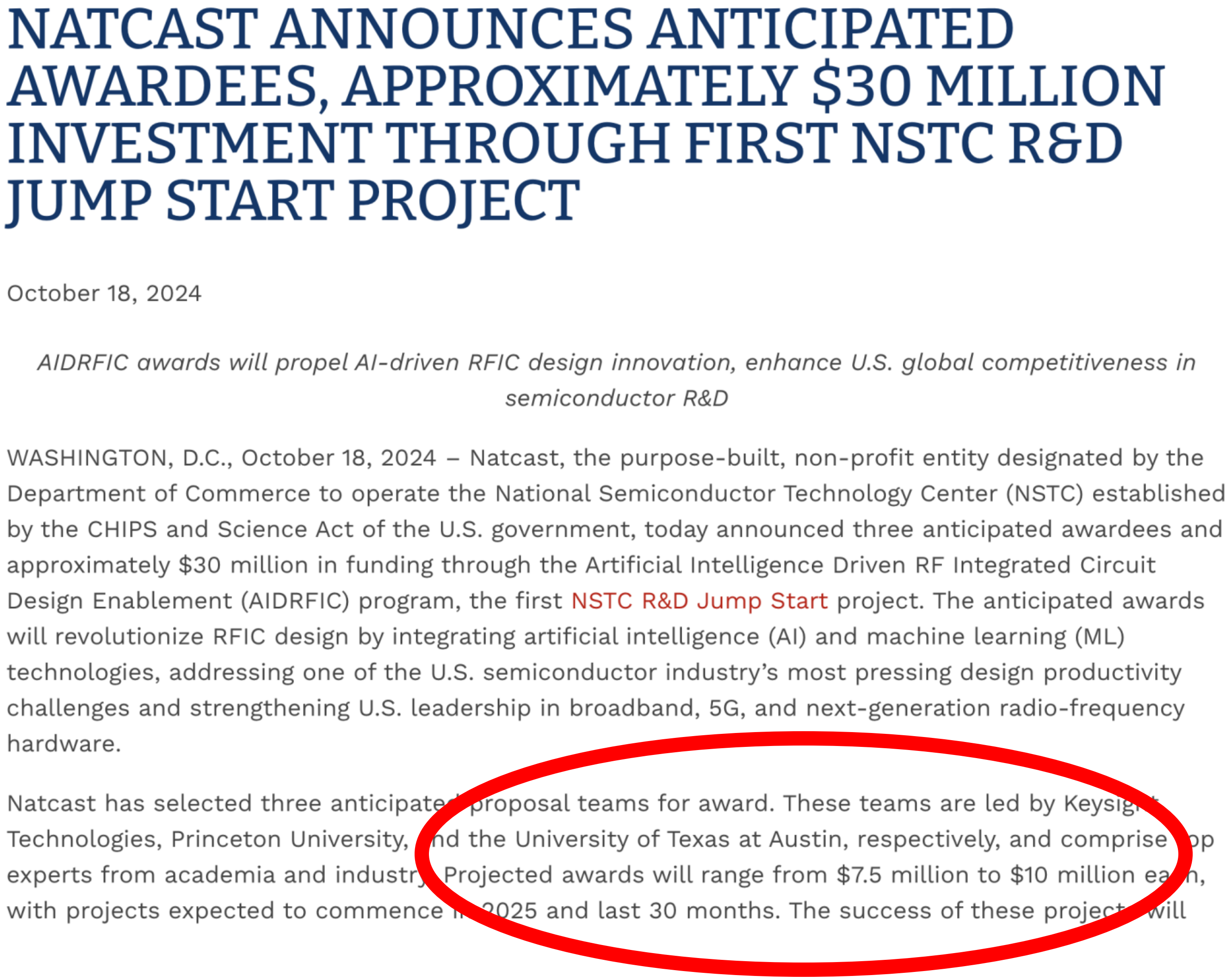
- ☹ Slow simulation restricts the number of optimization iterations possible
- ☹ Optimization is confined to a limited set of topology templates

- ☺ Physics-augmented ML model for fast design evaluation
- ☺ Bayesian optimization-based inverse design
- ☺ Super-human, non-intuitive designs

# 1st NSTC Jump Start R&D Program: AIDRFIC

Natcast

- **Active**: leverage analog DA
- **Passive:** PulseRF++
- Just scratched the surface!
- **75+ team competed => 3 winning teams**
- **UT Austin team** "GENIE-RFIC: Generative ENgine for Intelligent and Expedited RFIC Design"



## NATCAST ANNOUNCES ANTICIPATED AWARDEES, APPROXIMATELY $30 MILLION INVESTMENT THROUGH FIRST NSTC R&D JUMP START PROJECT

October 18, 2024

*AIDRFIC awards will propel AI-driven RFIC design innovation, enhance U.S. global competitiveness in semiconductor R&D*

WASHINGTON, D.C., October 18, 2024 – Natcast, the purpose-built, non-profit entity designated by the Department of Commerce to operate the National Semiconductor Technology Center (NSTC) established by the CHIPS and Science Act of the U.S. government, today announced three anticipated awardees and approximately $30 million in funding through the Artificial Intelligence Driven RF Integrated Circuit Design Enablement (AIDRFIC) program, the first NSTC R&D Jump Start project. The anticipated awards will revolutionize RFIC design by integrating artificial intelligence (AI) and machine learning (ML) technologies, addressing one of the U.S. semiconductor industry's most pressing design productivity challenges and strengthening U.S. leadership in broadband, 5G, and next-generation radio-frequency hardware.

Natcast has selected three anticipated proposal teams for award. These teams are led by Keysight Technologies, Princeton University, and the University of Texas at Austin, respectively, and comprise top experts from academia and industry. Projected awards will range from $7.5 million to $10 million each, with projects expected to commence in 2025 and last 30 months. The success of these projects will

# Conclusion

- Traditional electronics scaling cannot race with AI Model scaling
  - Emerging devices such as photonics for ML hardware
- Break the tradition that ML first and then hardware/system
  - Co-design/hardware-aware AI can unlock huge efficiency potential
- Hardware/chip design itself, e.g., modeling and LLM aided design
  - But still far away from super-human GenAI “all at once!” for chip design