Lightening-Transformer: A Dynamically-operated Optically-interconnected Photonic Transformer Accelerator

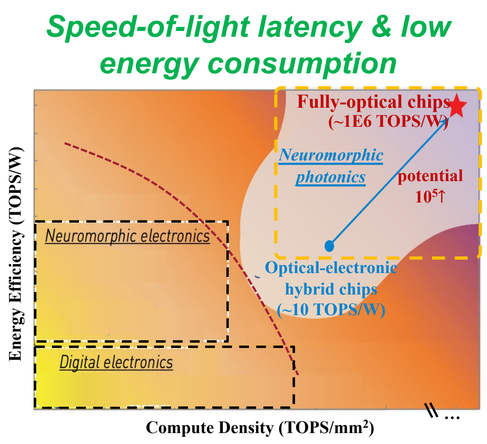
Hanqing Zhu[1], Jiaqi Gu[1,3], Hanrui Wang[2], Zixuan Jiang[1], Zhekai Zhang[2], Rongxing Tang[1], Chenghao Feng[1], Song Han[2], Ray T. Chen[1] and David Z. Pan[1]

[1]University of Texas at Austin, [2]Massachusetts Institute of Technology, [3]Arizona State University

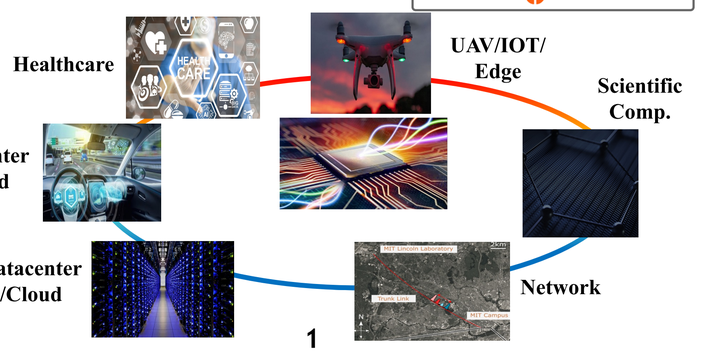
HPCA 2024

Photonic AI

Speed-of-light latency & low energy consumption



LIGHTMATTER, OpenLight, LIGHTELLIGENCE, OPTELLIGENCE, Lighton, CogniFiber, XANADU, LUMINOUS, OPTIUS, SYNOPSYS PHOTONIC SOLUTIONS, cadence PHOTONICS, Ansys LUMERICAL, amf ADVANCED MICRO FOUNDRY, AIM photonics, GlobalFoundries



1

Photonics for Transformer

Transformers own different MatMul features

- ***Dynamic MatMul***

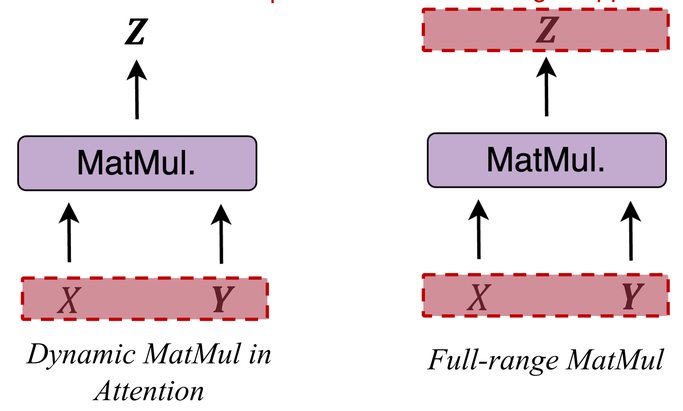
X, Y: Dynamically activations

Need Real-time and Freq. Switch

- ***Full-range MatMul***

X, Y, Z*: Full-range*

Need full-range support



Dynamic MatMul in Attention | *Full-range MatMul*

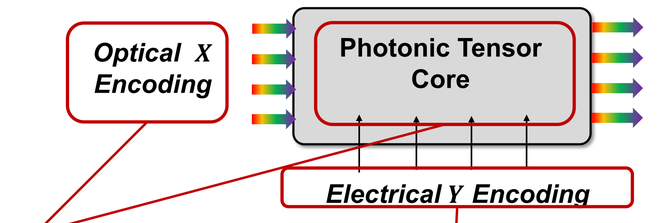
CNNs: a typical workload that photonics support

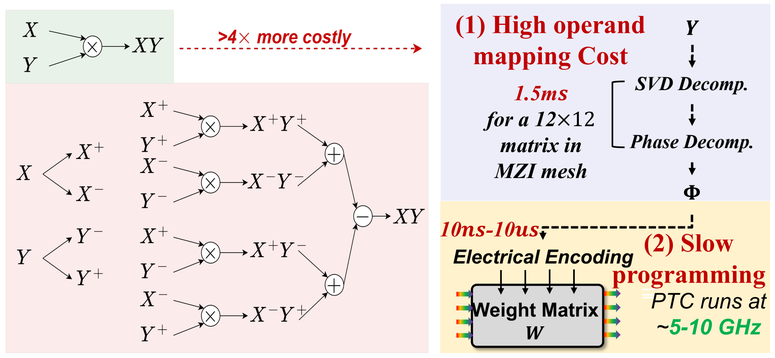
- *Weight-static MatMul*
- *Non-negative activation*

Existing Photonic AI Accl. as plug-in-and-play solution for Transformer?

2

Existing Photonics for Transformer



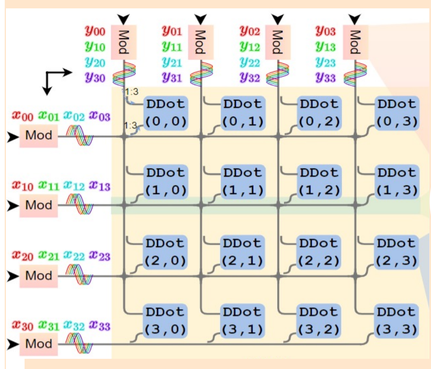


We propose the ***first energy-Efficient, high-performance*** Photonic Transformer Accelerator w/o the above issues!

3

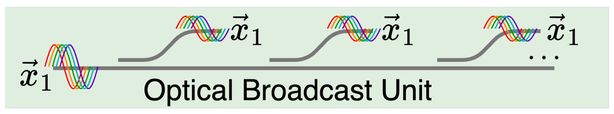
Dynamically-operated Photonic Tensor Core: A versatile photonic GEMM primitive

Versatile PTC Design

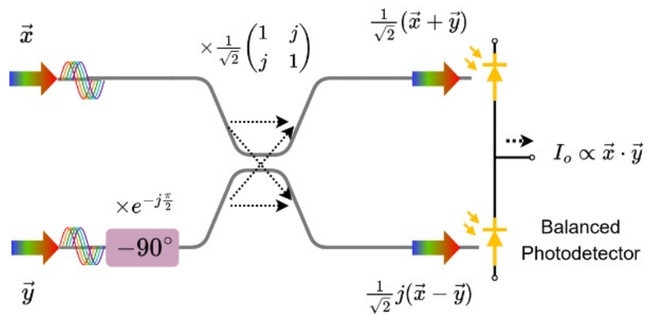


General full-range MatMul acceleration

- ***Ultra-parallel***: One-shot MatMul
  - Spectral and Spatial prallesim
- ***Energy-efficient D/A and E/O***:
  - Maximized operand sharing with a crossbar-structure



- ***Ultra-compact*** and ***fully-passive*** dot-product engine
  - 1 phase shifter + 1 directional coupler for dot-product
  - No active devices inside



- ***Light-light interaction***
  - All operands are high-speed optical signals
- ***Versatile*** primitive for GEMM w/o any constraints
- ***Superior inherent noise robustness***

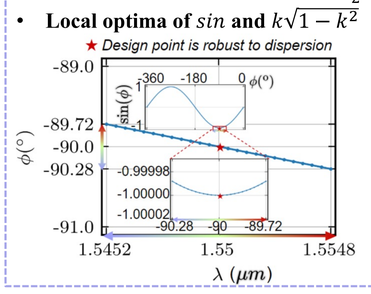
Noisy photonic dot-product w/ encoding and WDM dispersion

$$I_o \propto \sum_{i=0}^{N-1}\left(\frac{(k_i^2-t_i^2)\hat{x}_i^2+(t_i^2-k_i^2)\hat{y}_i^2}{2}+2t_ik_i\hat{x}_i\hat{y}_i(-\sin\phi_i)\right)$$

$$\propto \sum_{i=0}^{N-1}\left(\underbrace{2k_i\sqrt{1-k_i^2}(-\sin\phi_i)}_{\text{❶multiplicative noise}}\hat{x}_i\hat{y}_i+\underbrace{(2k_i^2-1)\frac{\hat{x}_i^2-\hat{y}_i^2}{2}}_{\text{❷additive noise}}\right),$$

Minimum sensitivity with perturbation

- Design point $\phi_i=-90°$ and $k=\frac{1}{2}$
- Local optima of sin and $k\sqrt{1-k^2}$



Self-compensation and scaling

- x/y in the same range $[-1, 1]$
- Square-effect and ½ scaling

4

Unique Arch-level Opt. Opportunity.: Explore the full power of Optics and Analog

Architecture-level Opt.

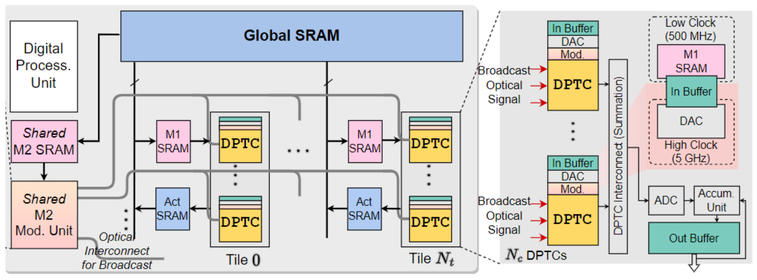
Explore photonic computing + photonic interconnect

And

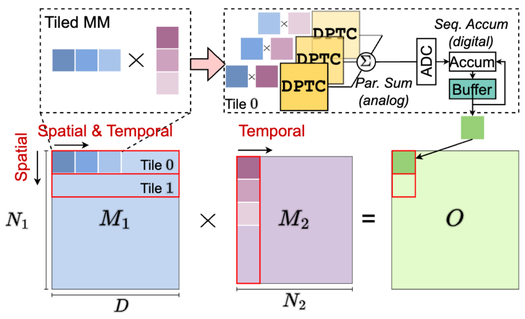
Explore Analog-domain data locality

Save signal conversion & data movement costs

A multi-core system

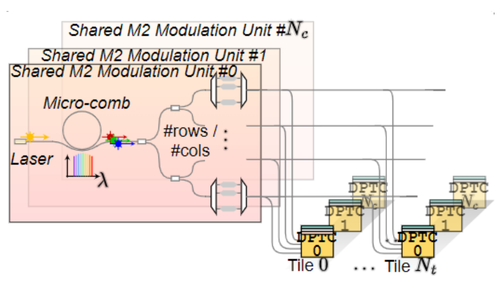


GEMM Mapping: OS dataflow (not supported in previous photonics)



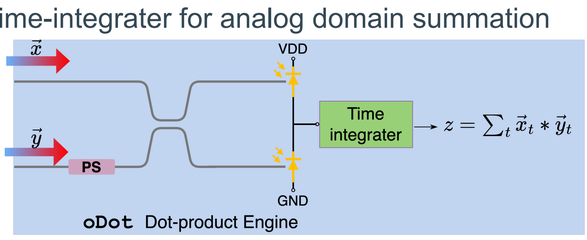
Opt 1: Photonic interconnect + Photonic computing

- *Low-cost long-distance* optical datamovement
- Broadcast shared M_2 data cross tile



Opt 2: Shift more workloads within analog domain

- Explore *analog locality* to reduce A/D cost
- Time-integrater for analog domain summation

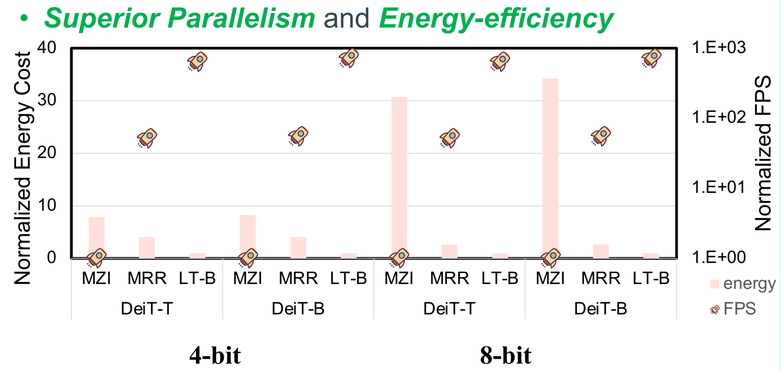


5

Experimental Results: A promising future of photonics for adv. ML tasks

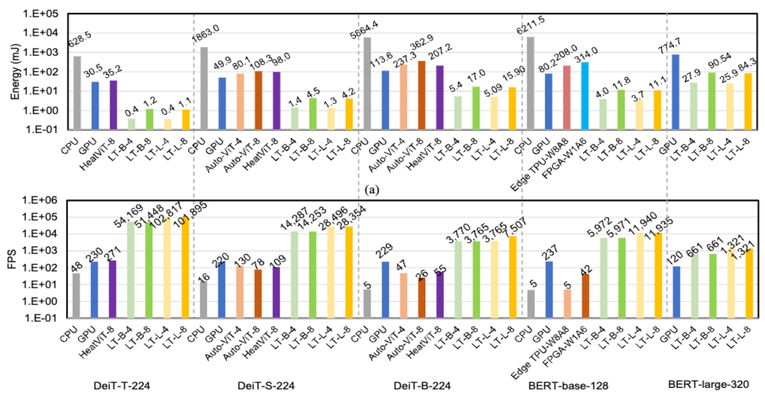
Compare with SOTA photonics

- $2.67-30.72\times$ more energy efficient
- $12-600\times$ speedup
- ***Superior Parallelism*** and ***Energy-efficiency***



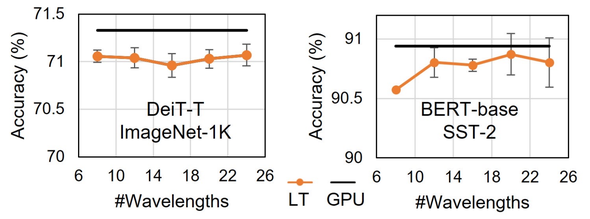
Compare with SOTA digitals

- $100-1000\times$ lower energy-delay product
- ***Huge potential of photonics***

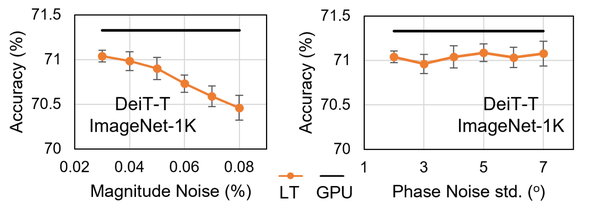


Noise-resilient analog computing core

- Inherent **robustness** to WDM dispersion
- Feature the high **scaling** ability with # wavelength



- $<0.5\%$ accuracy drop when varing variations



The Future of *Light-AI Interaction* is Bright

6

AE GitHub Repo

Contact: Hanqing Zhu (hqzhu@utexas.edu)

Acknowledgment:
Supported by AFOSR #FA9550-17-17-1-0071 and AFSOR #FA9550-23-1-0452

7