SESSION 2: MEMORY WALL AND INTERCONNECT WALL

Optical Interconnect for Scalable AI Systems

Seongwon (Gabriel) Yoon, PI: Prof. Shimeng Yu
gabrielyoon@gatech.edu

Georgia Tech School of Electrical and Computer Engineering

Agenda

OUTLINE OF THIS TALK

GenAI LLM Era

Distributed Infrastructure for ML

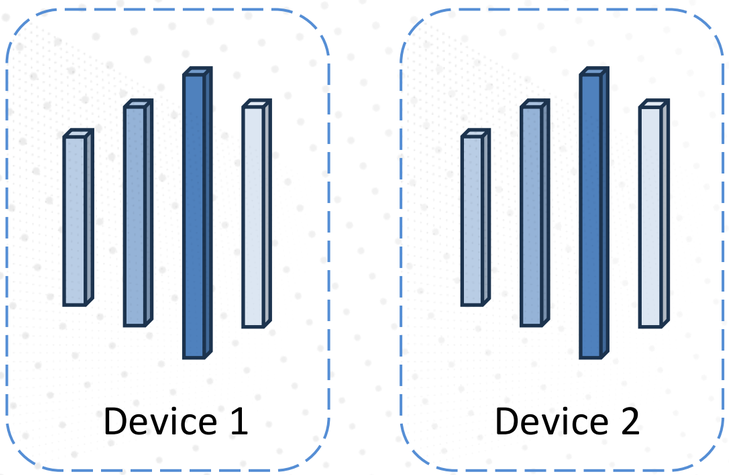
Optical Interconnect Solutions

GenAI LLM Era

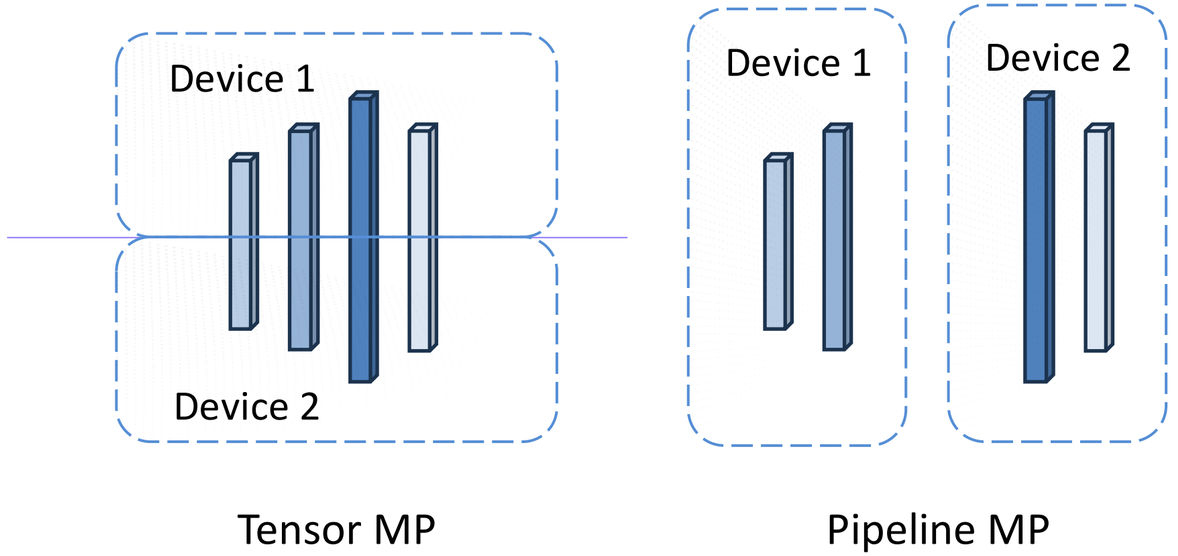
- Distributed Training
- LLM / Mixture-of-Experts
- All-to-All traffic pattern in MoE

Distributed Training: Tensor & Model Parallelism

Data Parallelism (DP)



Model Parallelism (MP)



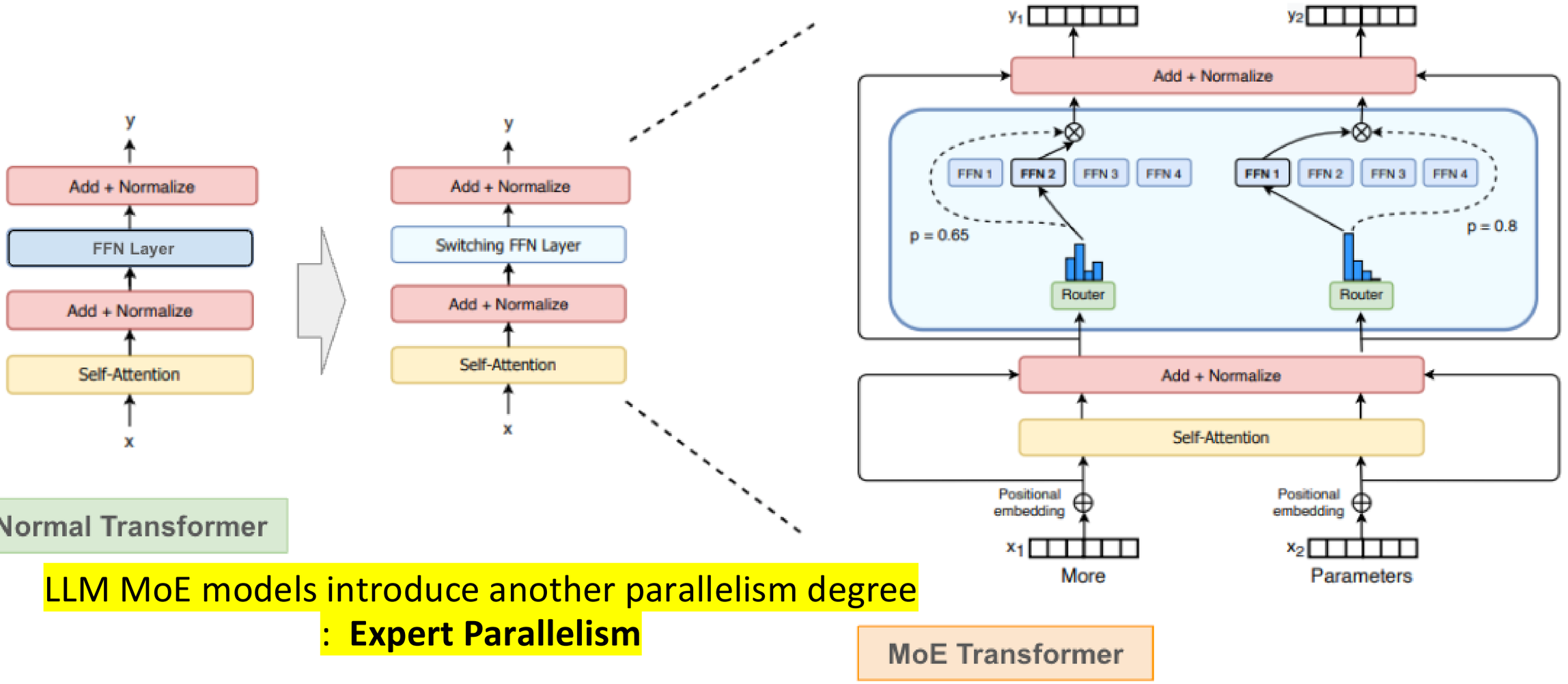
Tensor MP

Pipeline MP

Models and datasets must be sub-divided to fit into GPU memory for distributed training

Figure adopted from ECE 8803 HML Lecture Note by Prof. Divya Mahajan

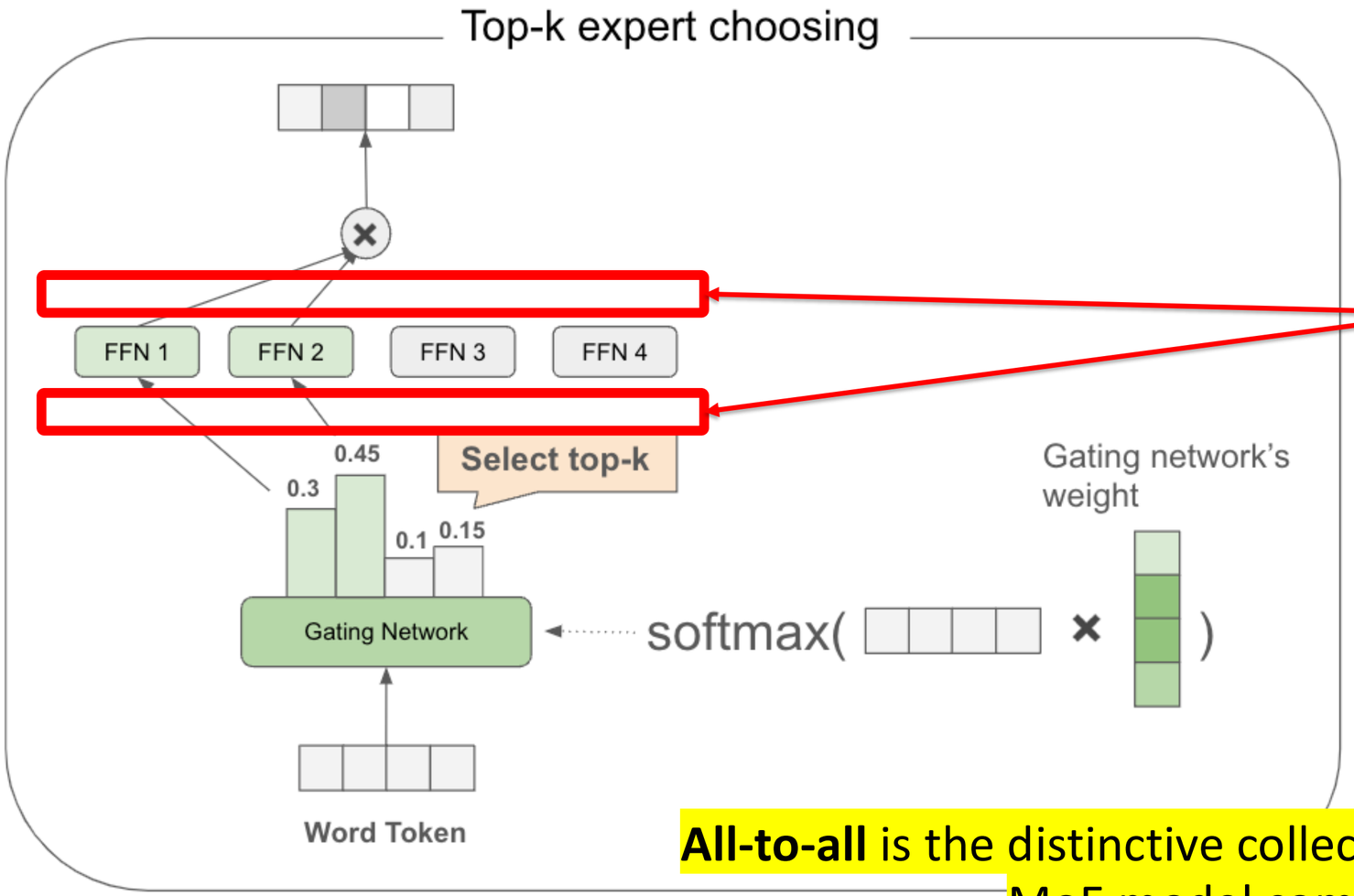
LLM Mixture-of-Experts (MoE)

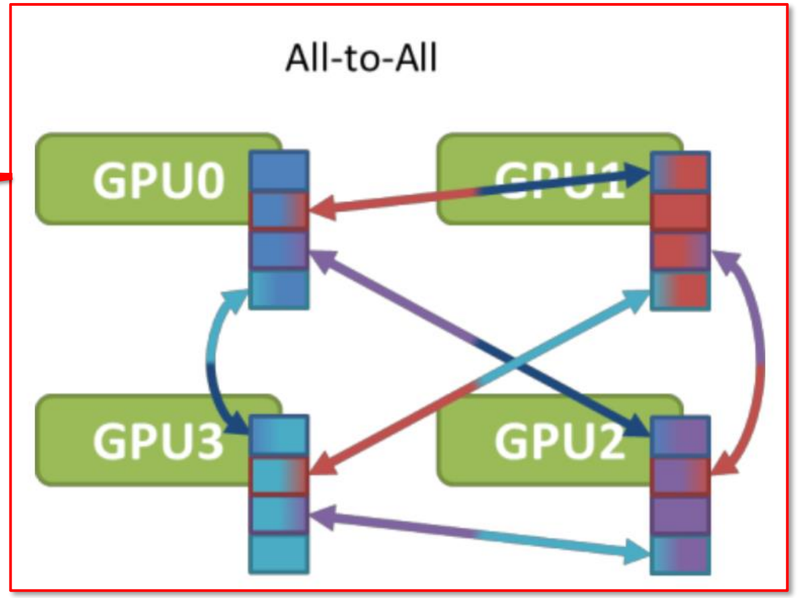


LLM MoE models introduce another parallelism degree : **Expert Parallelism**

Figure adopted from Zhongkai Yu et al., ”Orders in Chaos: Enhancing Large-Scale MoE LLM Serving with Data Movement Forecasting”, *arXiv*

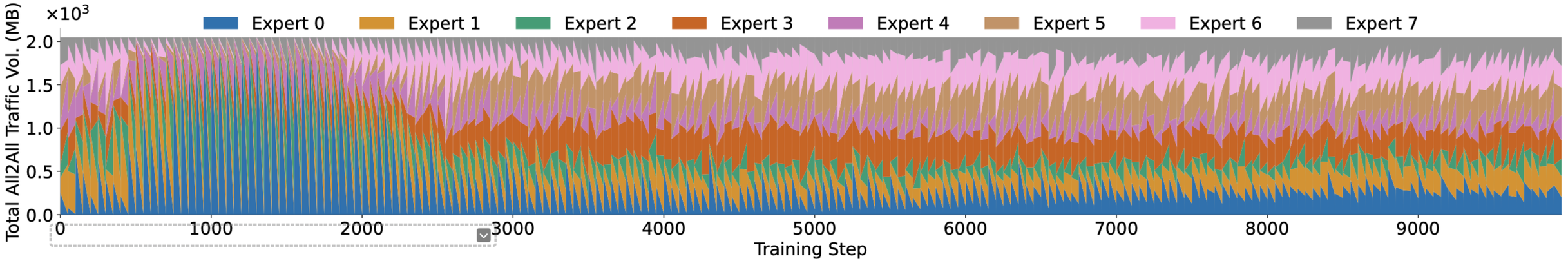
All-to-All traffic pattern in LLM MoE model



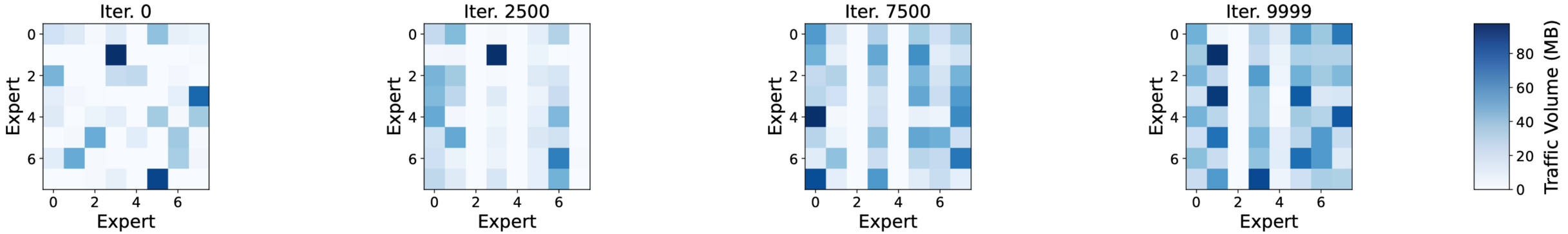


All-to-all is the distinctive collective communication pattern in LLM MoE model compared to dense LLM

All-to-All traffic pattern in LLM MoE model



(a) In the temporal dimension, the all-to-all traffic volume of each node varies across different training iterations.



(b) In spatial dimension, the all-to-all traffic volume is non-uniform across different experts.

Figure 4: [Mixtral 8×7B in production] All-to-all traffic dynamics during MoE training.

Expert selection shows strong spatial locality which makes a **hotspot GPU**

Figure adopted from Xudong Liao et al., “Mixnet: A Runtime Reconfigurable Optical-Electrical Fabric for Distributed Mixture-of-Experts Training”, ACM/IEEE SIGCOMM (2025)

Distributed Infrastructure for ML

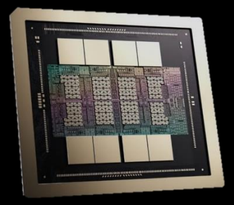
ML is changing datacenters, but how?

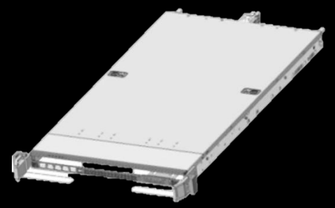
- Under-the-radar Datacenter Infrastructure
- Communication Hierarchy

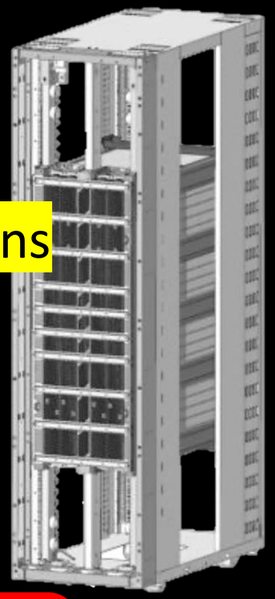
Georgia Tech School of Electrical and Computer Engineering

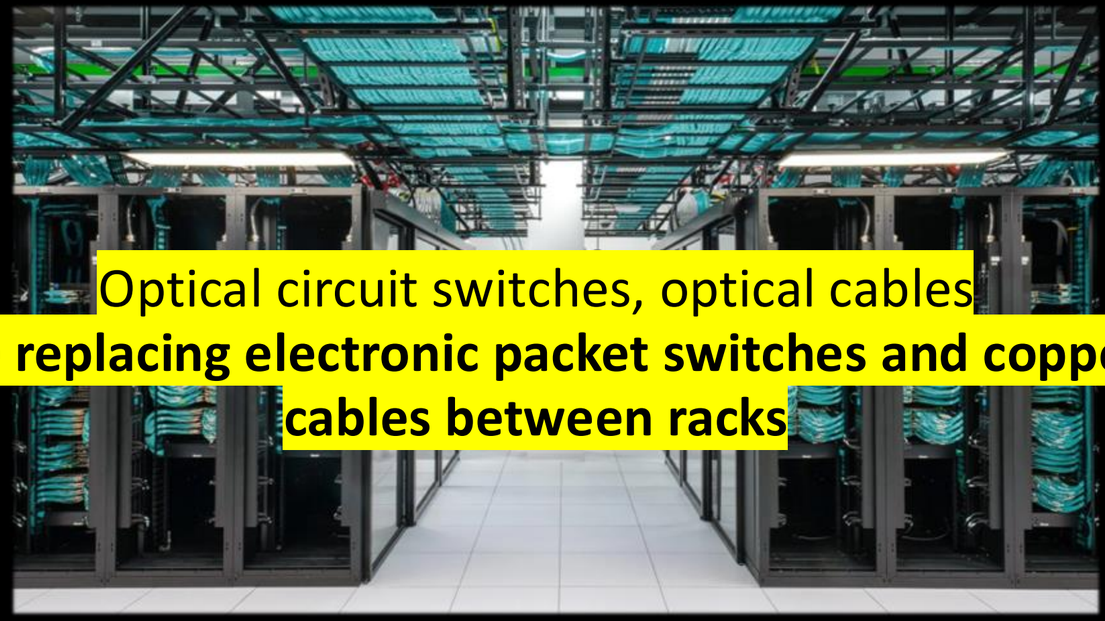
Distributed Infrastructure for ML

How to introduce optical connections **between chips/trays?**









Optical circuit switches, optical cables **are replacing electronic packet switches and copper cables between racks**

Chip — Package

Tray — Between Trays — Scale-Up

Rack — Between Racks — Scale-Out

Data Center — Between Data Center — Scale-Across

Figure adopted from HOTCHIPS 2025 AI Rack Tutorial, Gilad Shainer (Nvidia) “Co-Packaged Silicon Photonics Switches for Gigawatt AI Factories”

Communication Hierarchy

Taxonomy of Interconnects for AI Accelerators in Datacenter

Scale-Across

Between

Datacenters

How to introduce optical connections **between chips/trays?**

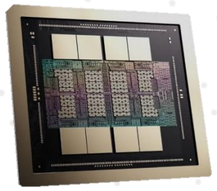
Scale-Up Domain

NVLink, UALink

Between GPUs and Switches

Inside a GPU

Package



Scale-Out Network

Ethernet or InfiniBand™

Connects accelerators in AI Cluster

Front Side Network

Ethernet

Connects everything in the datacenter (storage, compute tiers)

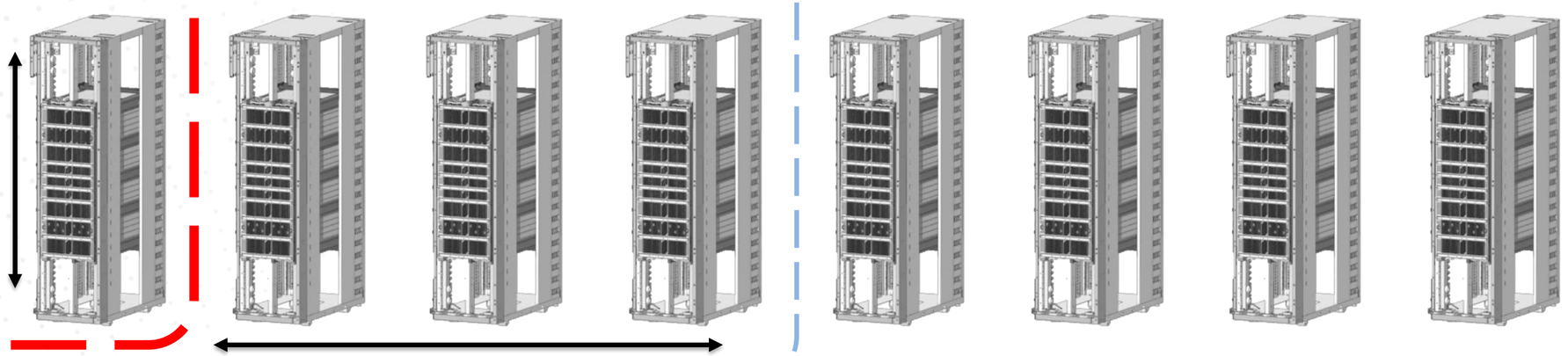


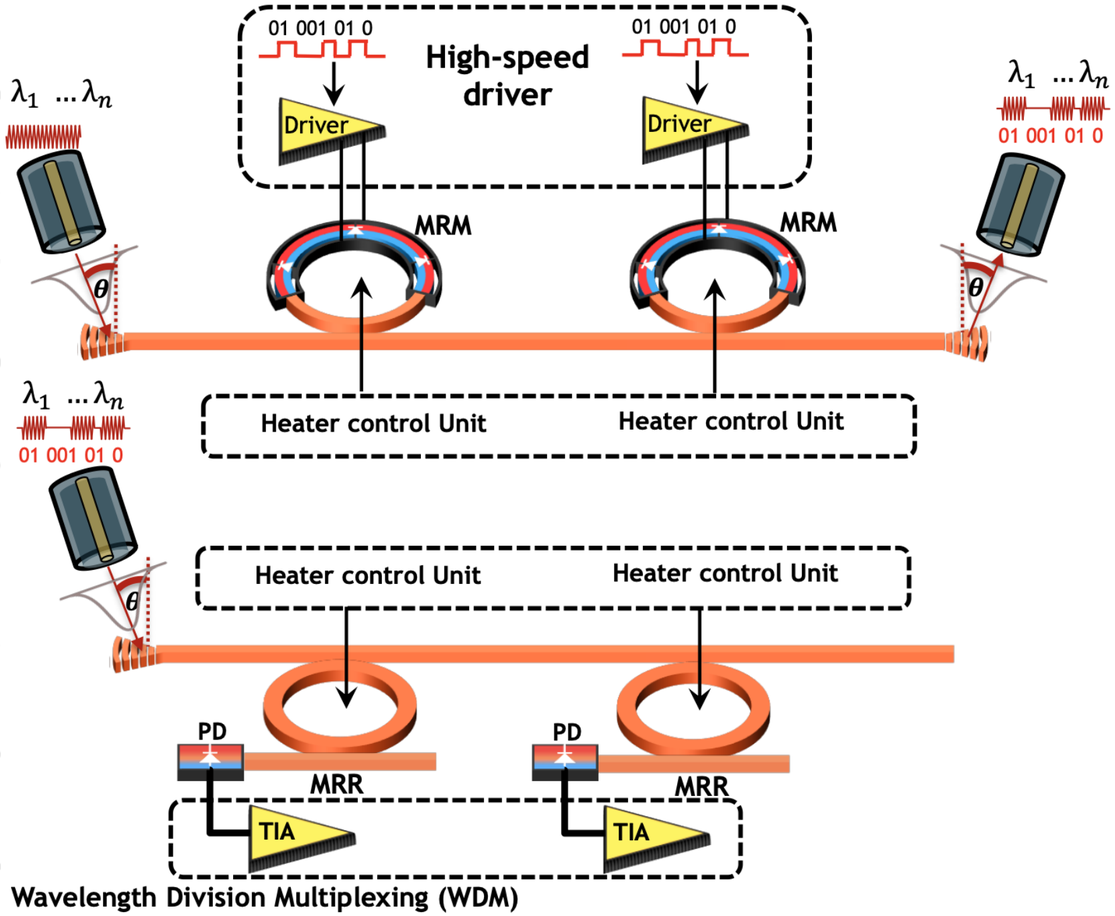
Figure adopted from HOTCHIPS 2025 AI Rack Tutorial, John Norton (Nvidia) “Case study : Nvidia GB200NVL72”

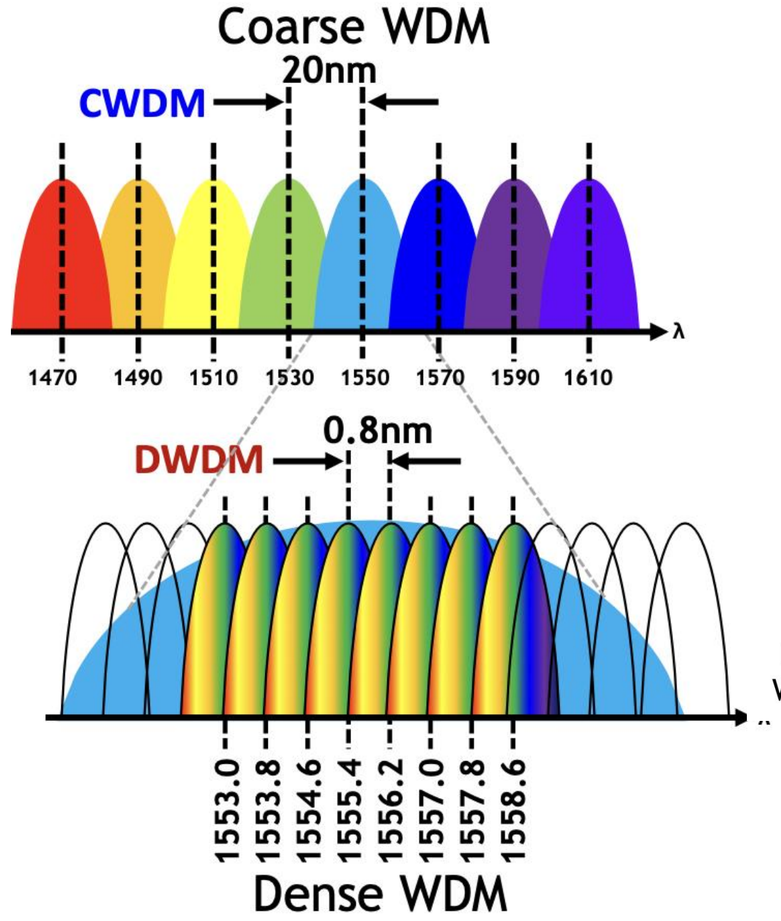
Optical Interconnect Solution in Scale-up Domain

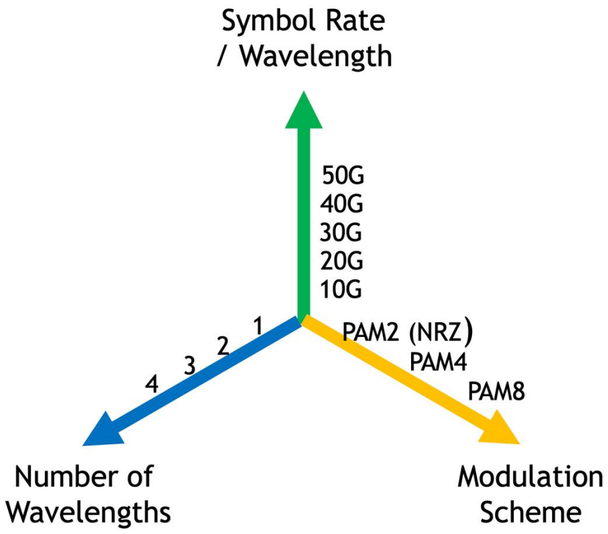
How to introduce optical module?

- Optical Device Module
- 1. Co-Packaged Optics (CPO)
- 2. Optical Interposer

Bandwidth Scaling with Optical Interconnect





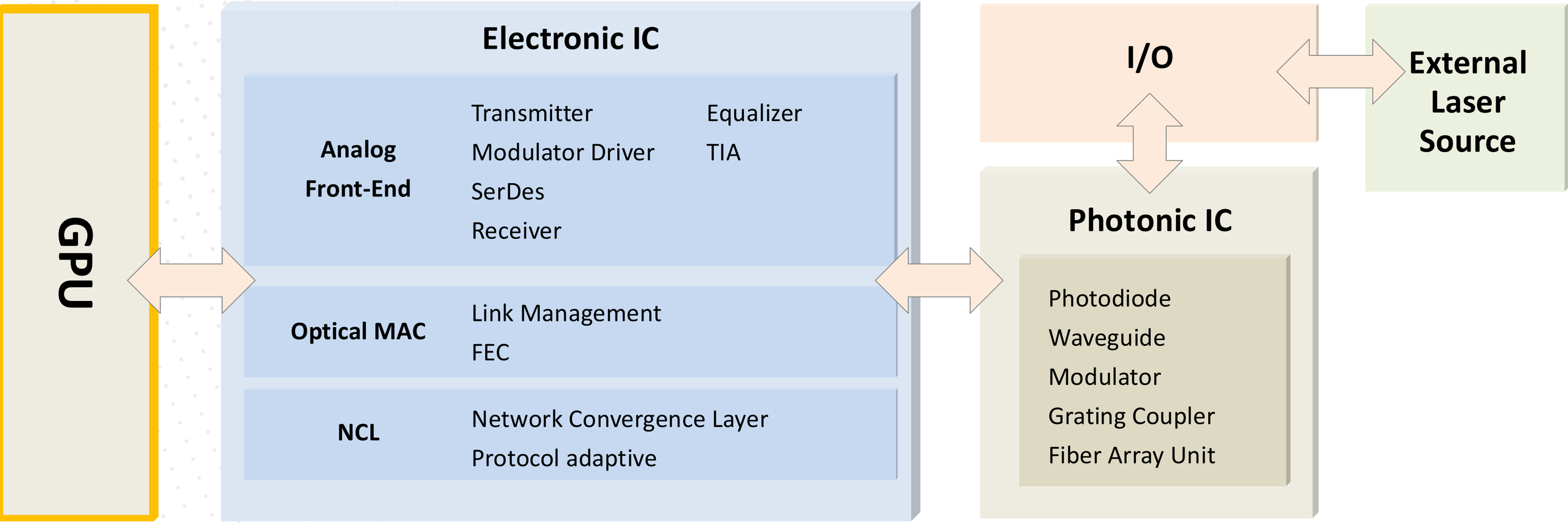


Wavelength Division Multiplexing enables bandwidth scaling

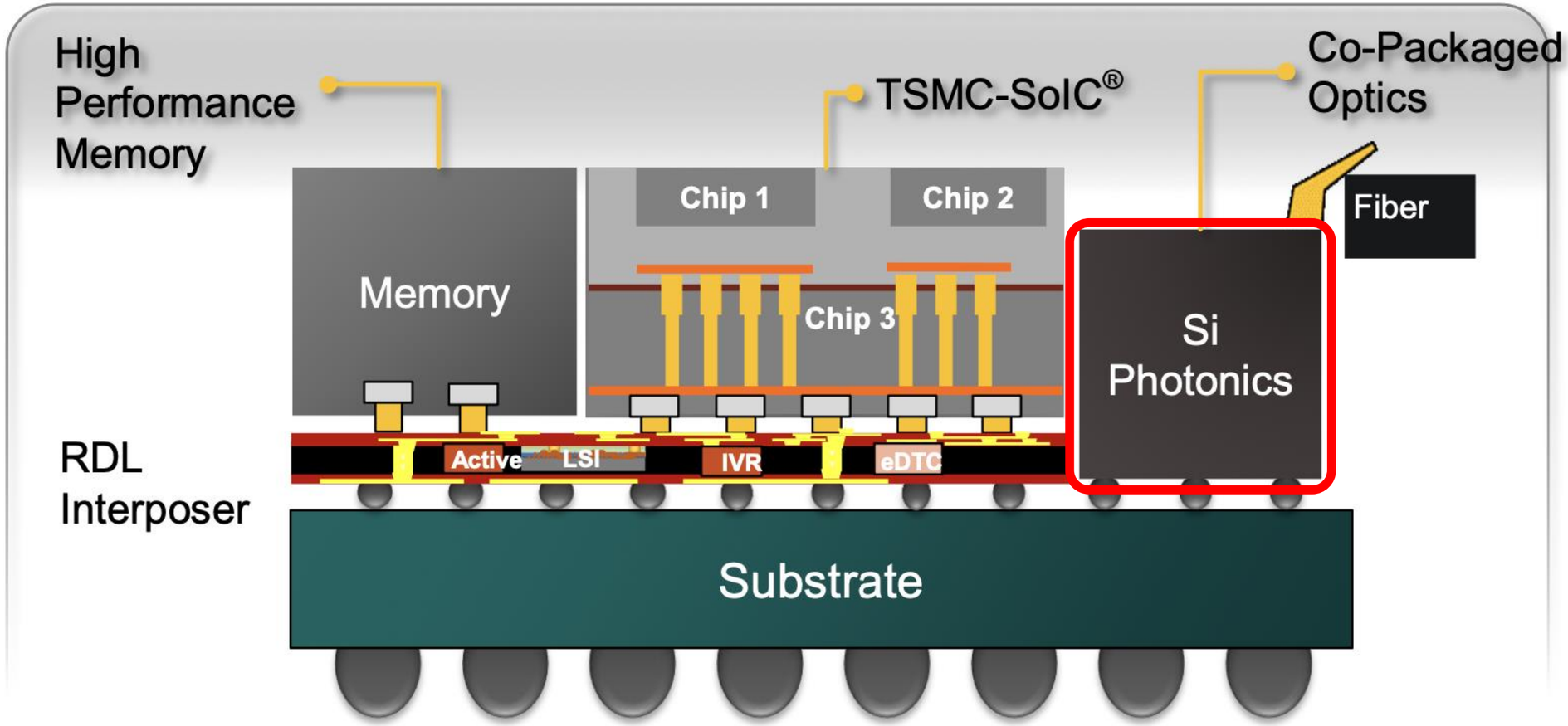
Figure adopted from John H. Lau et al., “Co-Packaged Optics—Heterogeneous Integration of Photonic Integrated Circuits and Electronic Integrated Circuits”

Optical Interconnect Overview

Optical connection requires electrical-optical conversion



TSMC COUPE platform for optical connection

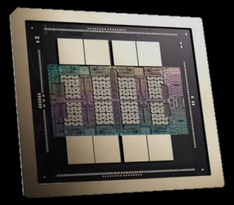


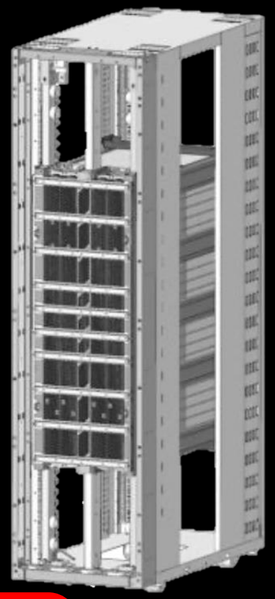
TSMC offers optical packaging platform with stacked EIC / PIC

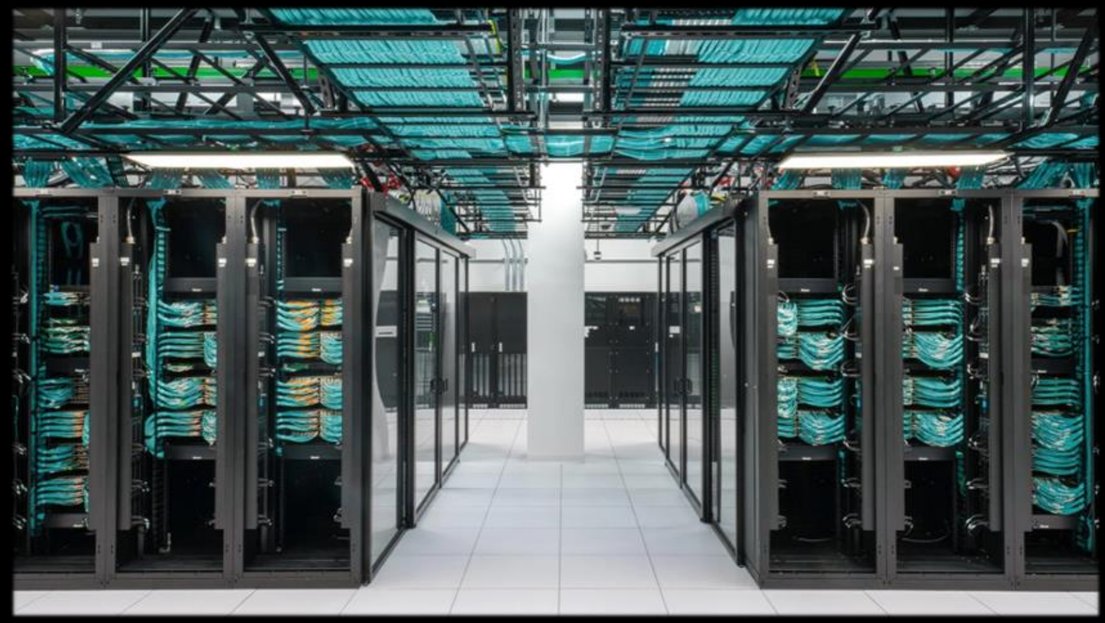
Figure adopted from John H. Lau et al., “Co-Packaged Optics—Heterogeneous Integration of Photonic Integrated Circuits and Electronic Integrated Circuits”

Distributed Infrastructure for ML

Let’s look at the two ways to **introduce optical interconnect**







Chip — Package

Tray — Between Trays — Scale-Up

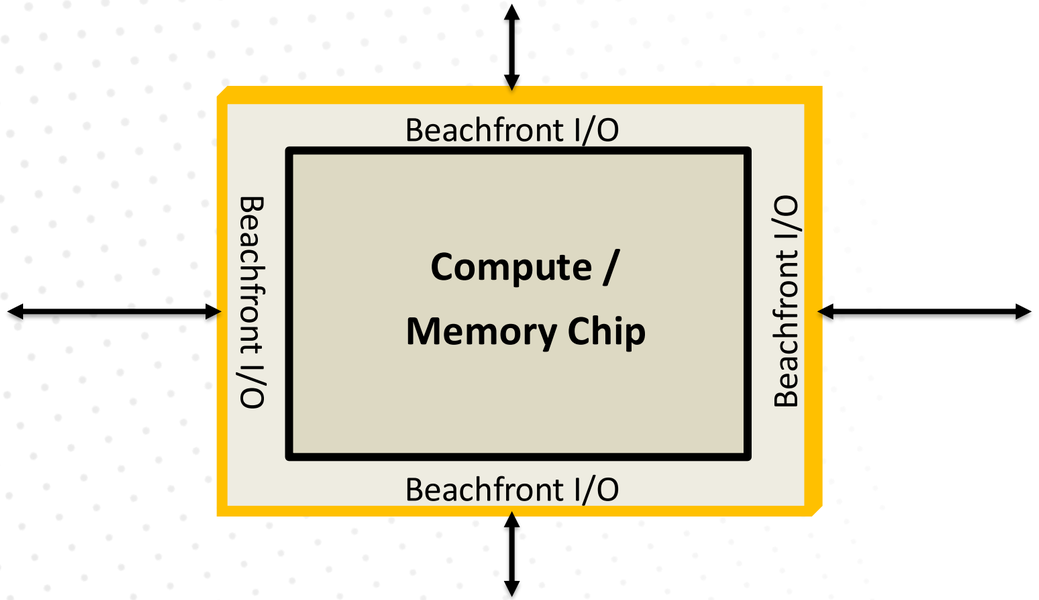
Rack — Between Racks — Scale-Out

Data Center — Between Data Center — Scale-Across

Figure adopted from HOTCHIPS 2025 AI Rack Tutorial, Gilad Shainer (Nvidia) “Co-Packaged Silicon Photonics Switches for Gigawatt AI Factories”

How to introduce Optical Module in Scale-up Domain?

Two options



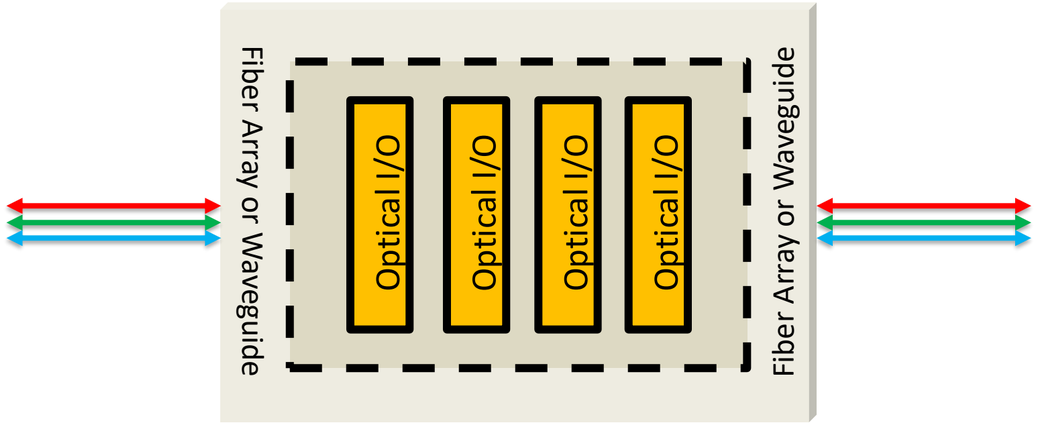
1. Electronic Fabric + CPO

PCIe, UCIe

Communication happens at the chip parameter

FLOPs ∝ Chip Area (L^2)

I/O Bandwidth ∝ Chip Perimeter (L)



2. Optical Interposer

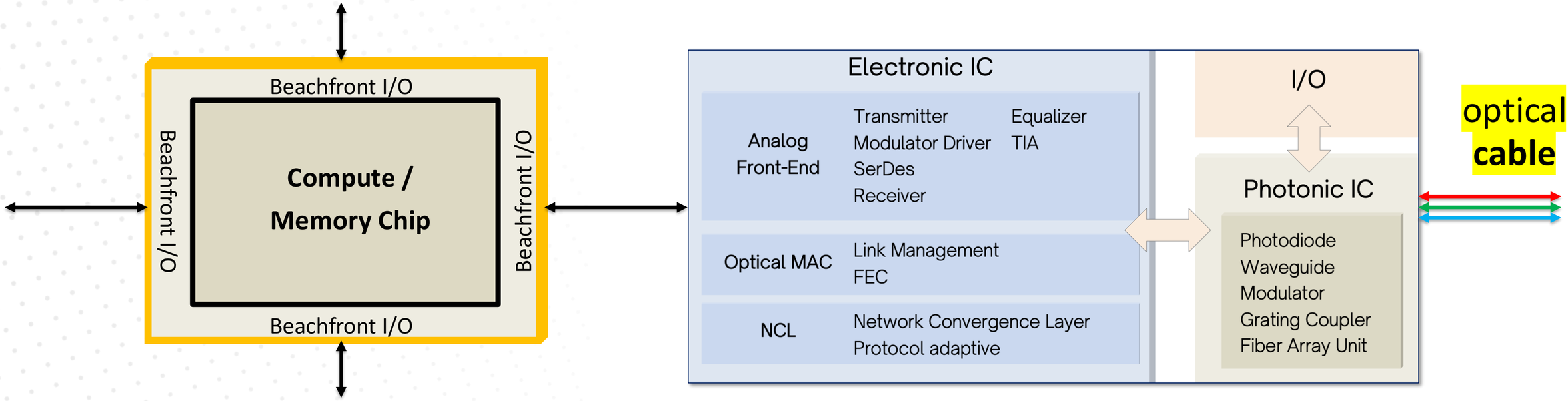
Silicon Photonics

I/O Bandwidth scales with compute area by 3D packaging

I/O Bandwidth ∝ Chip Area (L^2)

How to introduce Optical Module in Scale-up Domain?

1. Electronic fabric + co-packaged optics



1. Electronic Fabric + CPO

PCIe, UCIe

Communication happens at the chip parameter

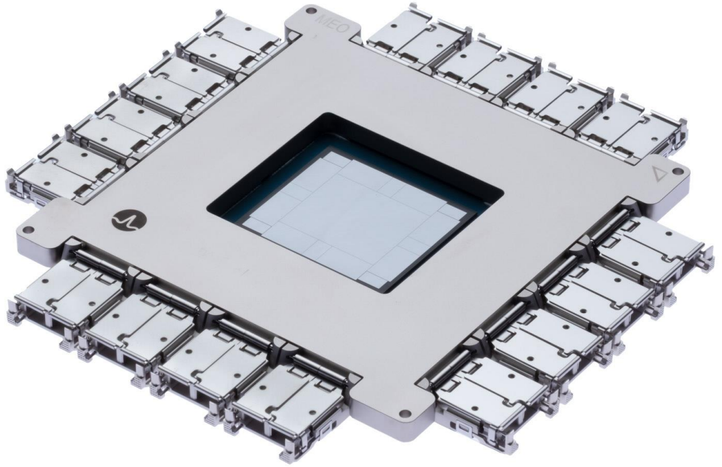
FLOPs ∝ Chip Area (L^2)

I/O Bandwidth ∝ Chip Perimeter (L)

First option :
conventional electronic fabric with
co-packaged optics

Co-Packaged Optics (CPO) in 2.5D Integration

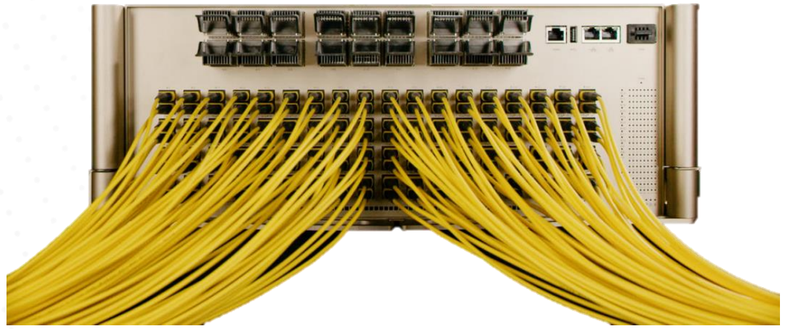
BROADCOM®



Tomahawk

Ethernet switch ASIC
102.4 Tbps

NVIDIA®



Quantum-X Photonics Switch

InfiniBand switch system / platform
144 ports x 800 Gb/s (115.2 Tbps)

AyarLabs

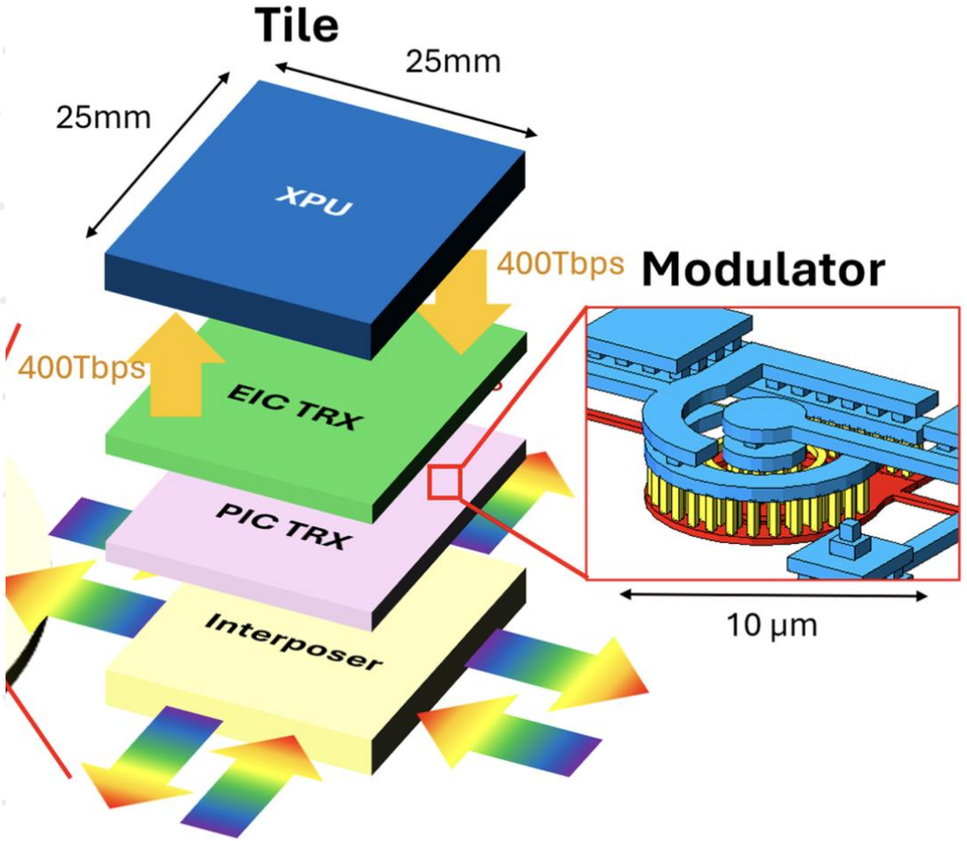
TeraPHY™ Optical Engine

UCIe-based optical I/O chiplets
8 Tbps per chiplet

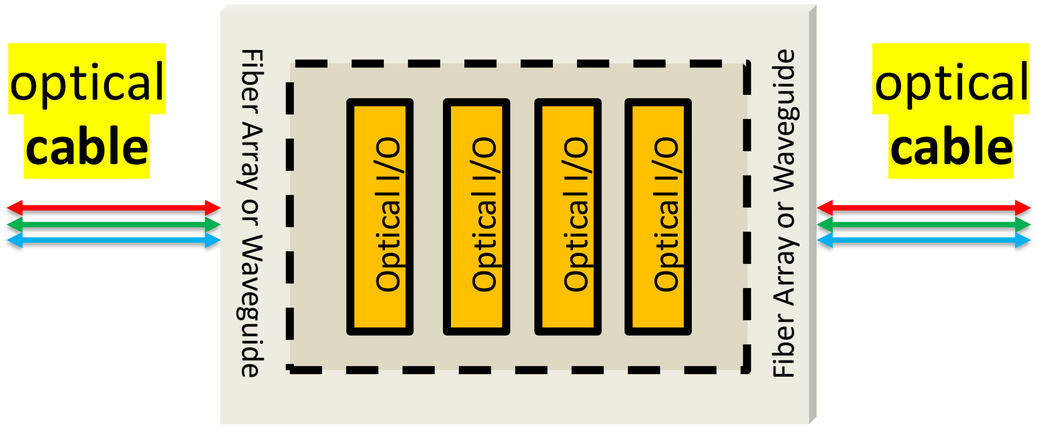
Figure adopted from HOTCHIPS 2025 AI Rack Tutorial, Gilad Shainer (Nvidia) “Co-Packaged Silicon Photonics Switches for Gigawatt AI Factories”

How to introduce Optical Module in Scale-up Domain?

2. Optical Interposer



Second option : waveguide based
GPU - EIC - PIC stack 3D Integration



2. Optical Interposer

Silicon Photonics

I/O Bandwidth scales with compute area
by 3D packaging

I/O Bandwidth $\propto$ Chip Area (L^2)

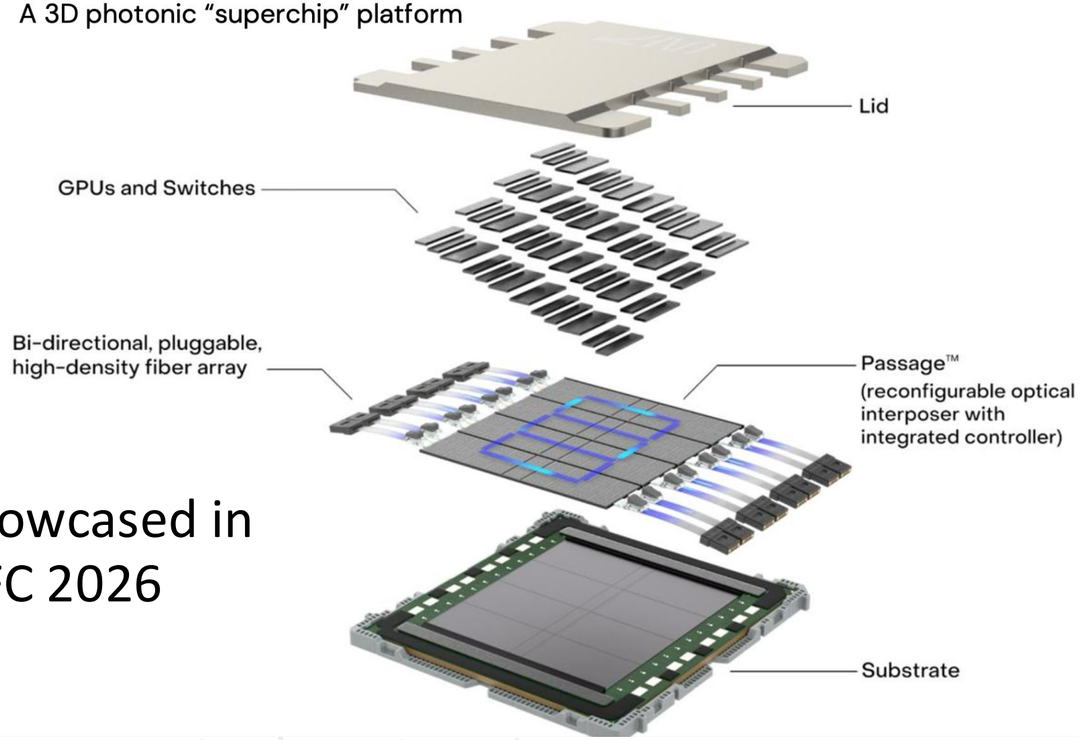
Optical Interposer with 3D integration

Top Layer : Compute and memory chiplets
Base Layer : lasers, waveguides, and optical switching and routing

LIGHTMATTER

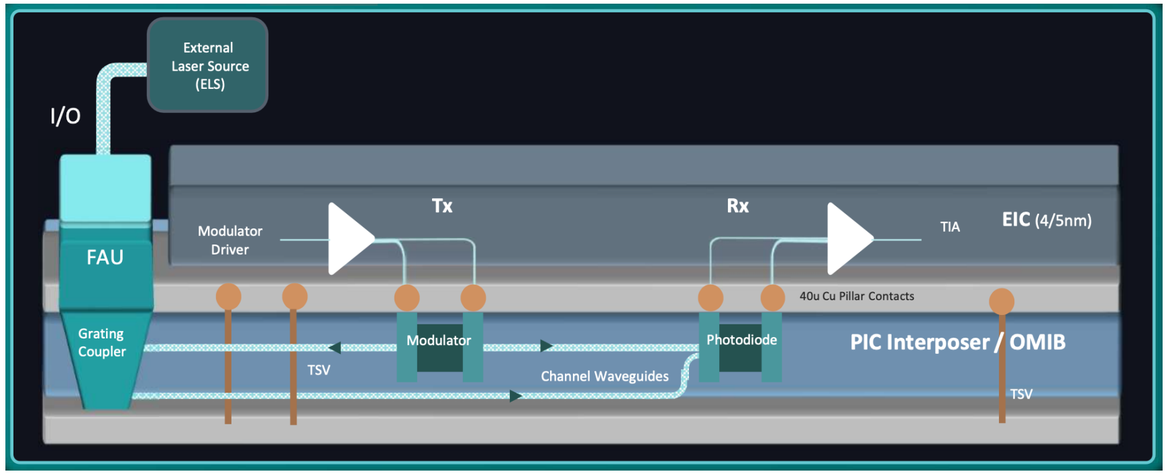
Passage M1000

A 3D photonic “superchip” platform



Showcased in OFC 2026

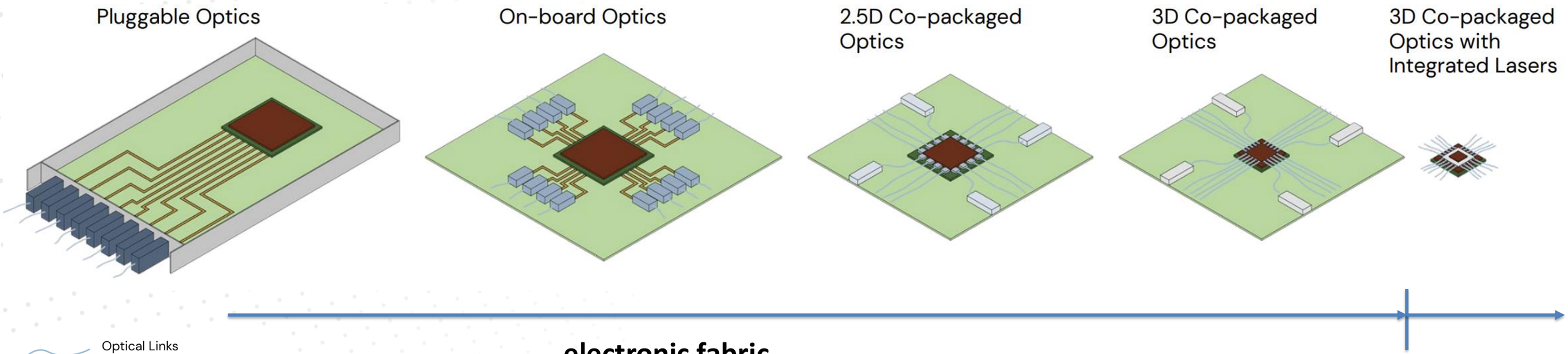
celestial AI™



Key advantage of optical interposer :
Delivers continuous 2D surface for optical I/O

Figure adopted from HOTCHIPS 2025 AI Rack Tutorial, Lightmatter, Celestial AI

Optical interconnect in chip- rack scale summary

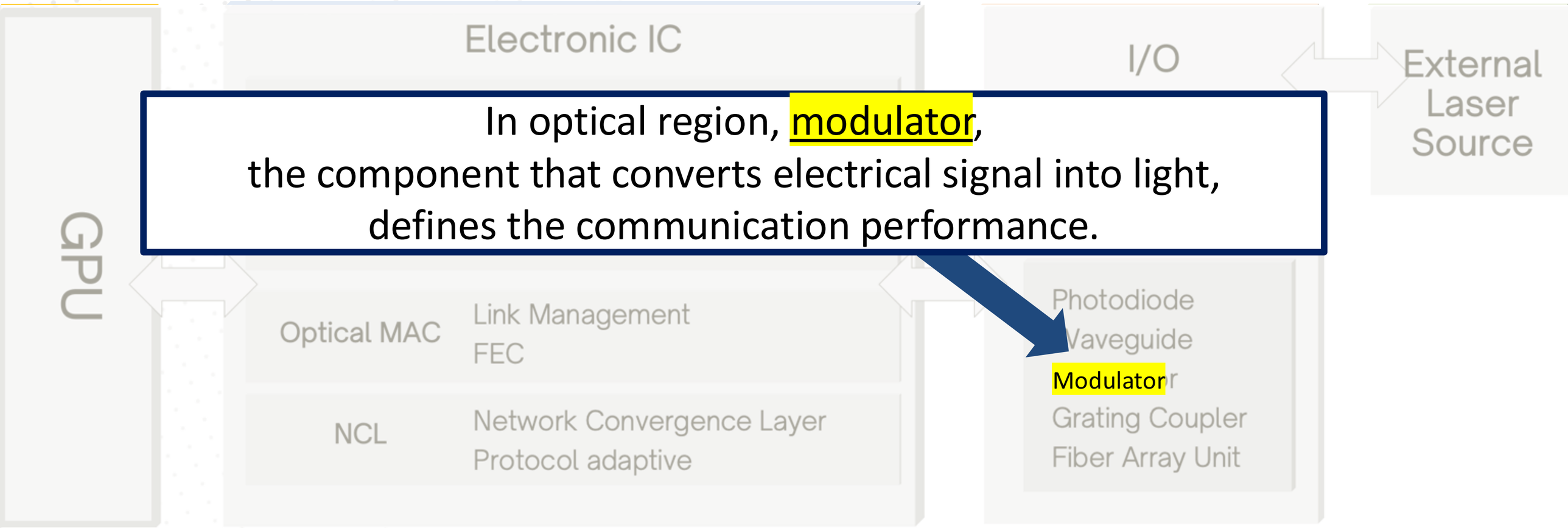


In short, optics are approaching closer and closer to the chip.
But what is the bottleneck in bringing optics closer to the processor?

Figure adopted from HOTCHIPS 2023 AI Rack Tutorial, Lightmatter, Celestial AI

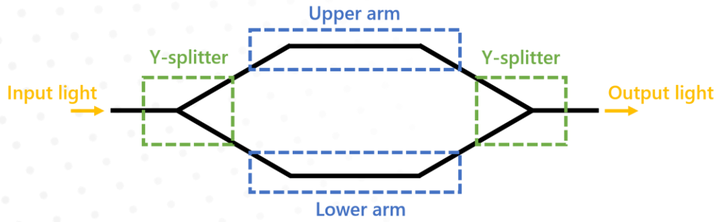
Recall Optical Interconnect System

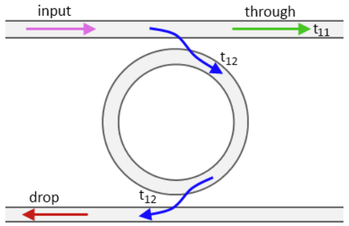
Optical connection requires electrical-optical conversion

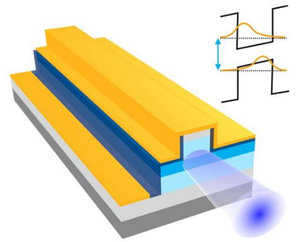


In optical region, modulator,
the component that converts electrical signal into light,
defines the communication performance.

Modulators inside PICs are based on silicon photonics



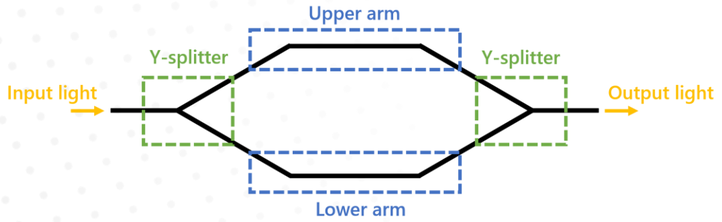


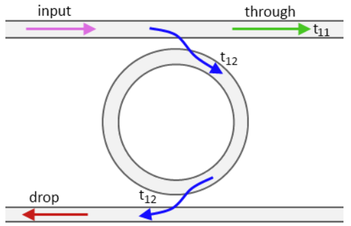


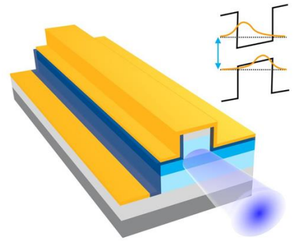
Metric	**Mach-Zehnder Interferometer (MZI)**	**Microring Resonator (MRR)** *Mostly thermo-optic*	**Electro-Absorption Modulator (EAM)**
Size	Large (1000 - 2000 um)	**Very small (< 50 um)**	Small (30 - 100 um)
Thermal Stability	Robust (Stable for over 50K fluctuation)	Sensitive (Must be stable under 1K range)	Good (Stable for over 50K fluctuation)
Speed (Data Rate)	< 56 Gbps	< 56 Gbps	< 56 Gbps
Power	High (require tens of voltages)	**Low**	Low
Optical Bandwidth	Broadband	< 1nm	~ 10s nm
WDM-friendly	Low (require MUX)	**High (resonance shift)**	Low (require MUX)
Operation Complexity	Low	High (resonance shift)	Low
XPU Integration	Low (high power)	Low (thermal crosstalk)	High

thermally-stable / electro-optic based

We propose Ferroelectric MRR

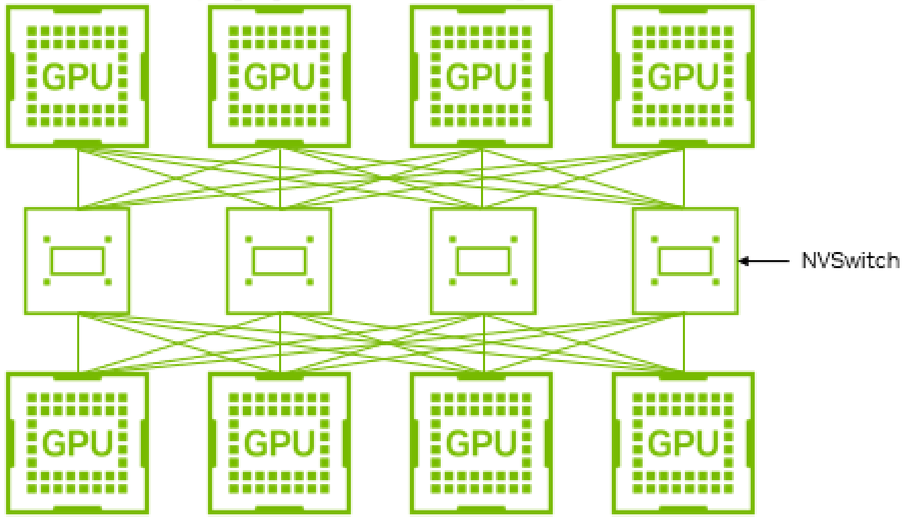




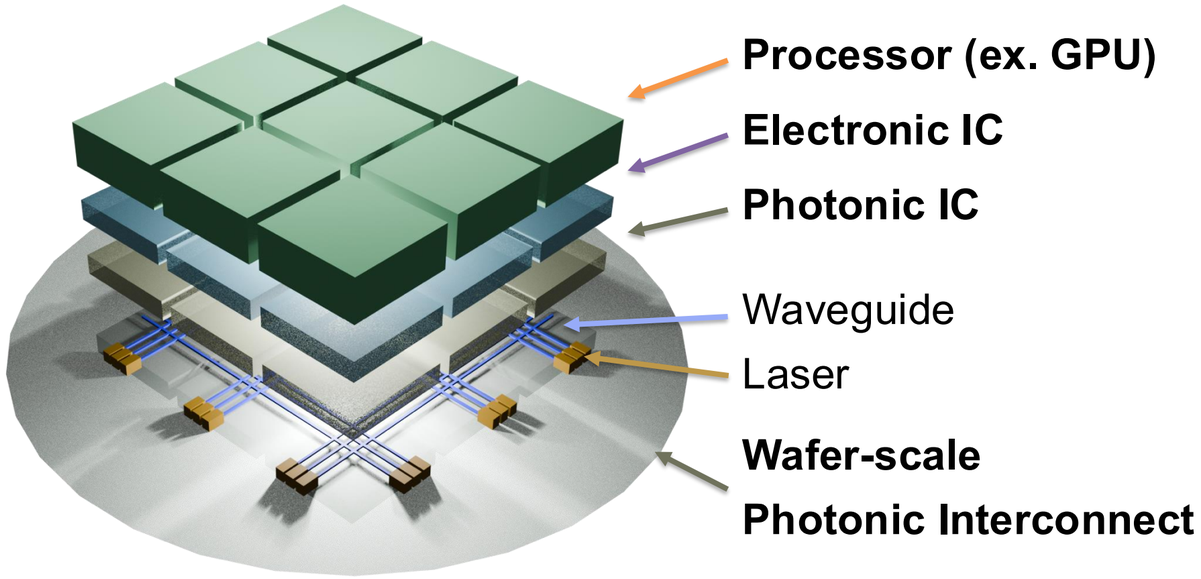


Metric	**Mach-Zehnder Interferometer (MZI)**	**Microring Resonator (MRR)** ~~*Mostly thermo-optic*~~	**Electro-Absorption Modulator (EAM)**
Size	Large (1000 - 2000 um)	**Very small (< 50 um)**	Small (30 - 100 um)
Thermal Stability	Robust (Stable for over 50K fluctuation)	~~Sensitive (Must be stable under 1K range)~~ Robust	Good (Stable for over 50K fluctuation)
Speed (Data Rate)	< 56 Gbps	< 56 Gbps	< 56 Gbps
Power	High (require tens of voltages)	**Low**	Low
Optical Bandwidth	Broadband	< 1nm	~ 10s nm
WDM-friendly	Low (require MUX)	**High (resonance shift)**	Low (require MUX)
Operation Complexity	Low	~~High (resonance shift)~~ Low	Low
XPU Integration	Low (high power)	~~Low (thermal crosstalk)~~ High	High

Wafer/Panel-Scale Computing Platform



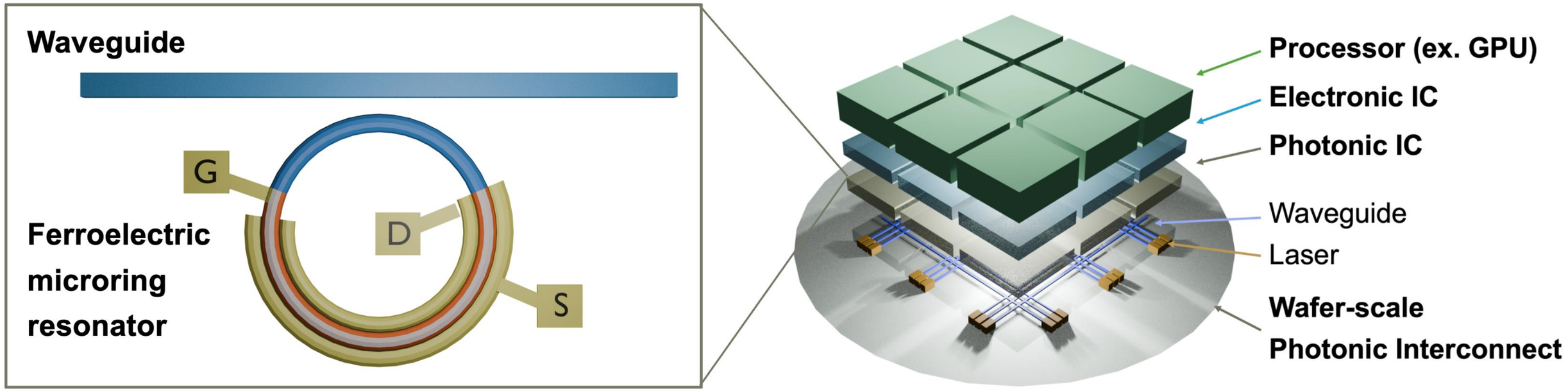
Nvidia DGX



Wafer-scale Platform

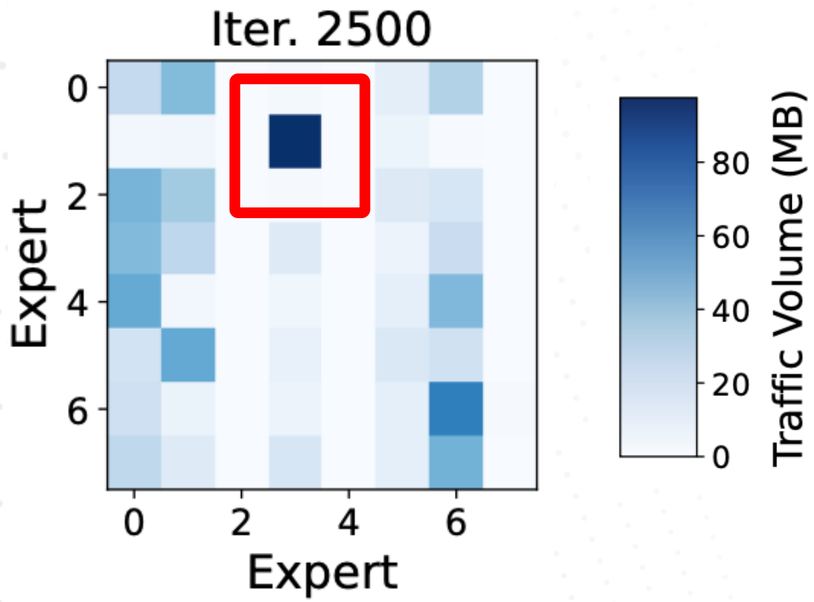
Wafer/panel-scale platform significantly increases the inter-GPU bandwidth compared to conventional multi-GPU node system.

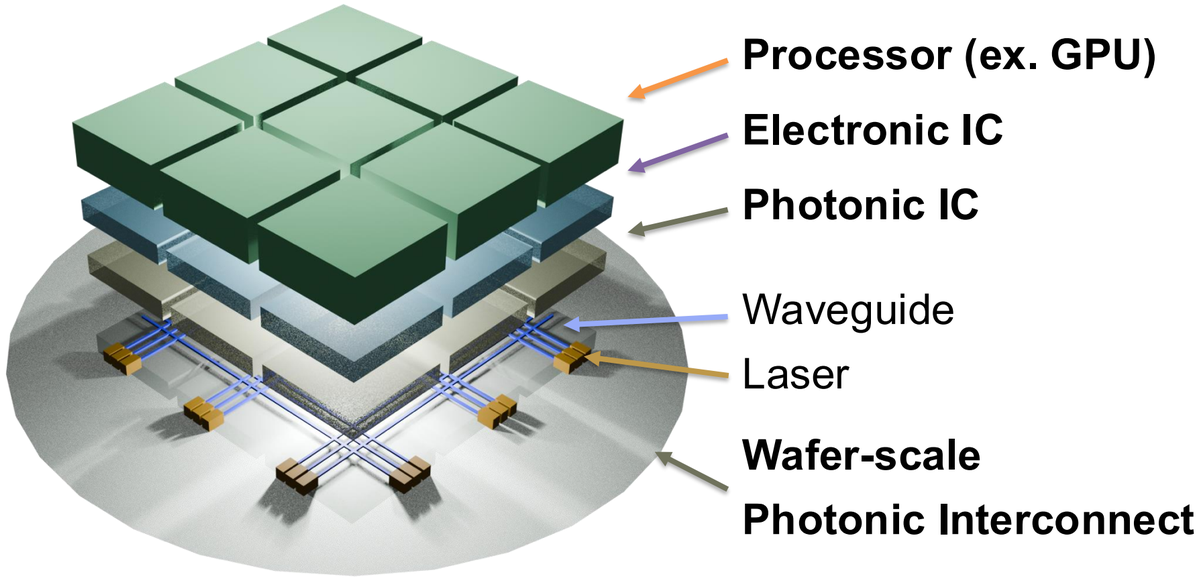
Ferroelectric MRR in wafer-scale optical interposer



However, thermal MRRs in the underlying PIC is prone to thermal fluctuation.

Ferroelectric MRR in wafer-scale optical interposer

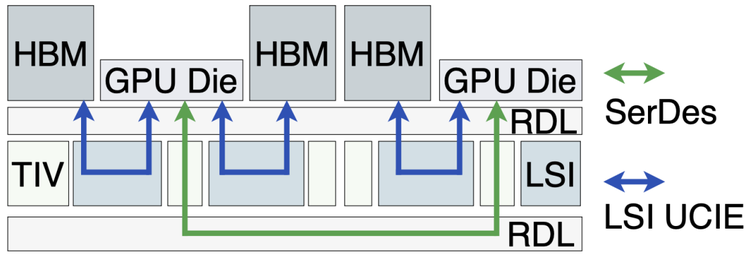
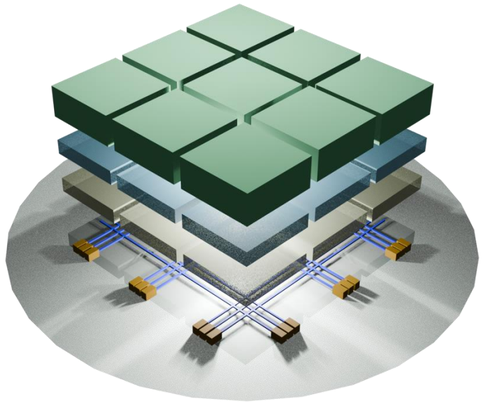




However, thermal MRRs in the underlying PIC is prone to thermal fluctuation.

We introduce ferroelectric MRRs simulated in LLM MoE model training scenario in which thermal hotspots are generated in **experts computation**

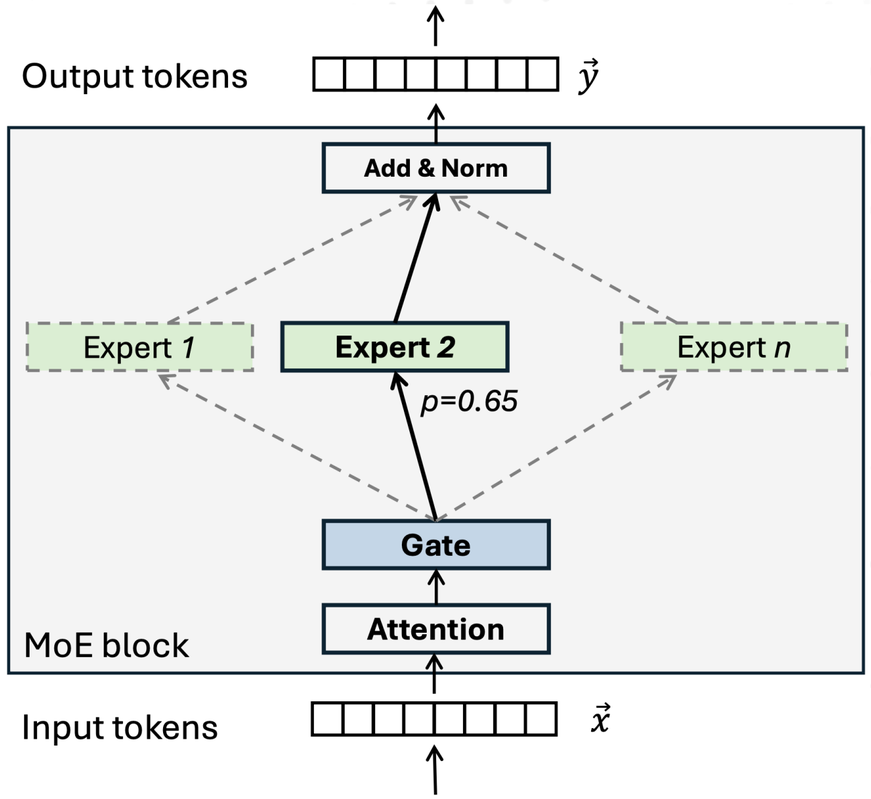
System-on-Wafer Architecture Comparison

	GPU	D2D Bandwidth	Hops	Latency	Structure
System -on-Wafer (electronic fabric)	4 x 4	1.5 TB/s	max 6	50 ns	
System -on-Wafer with optical interposer	4 x 4	1.5 TB/s	1 or 2	39 ns	

Thermal MRR

Ferroelectric MRR

Profile LLM MoE models



FlexFlow
DNN simulator

Add and Norm

All-to-all communication

Expert

All-to-all communication

Gate

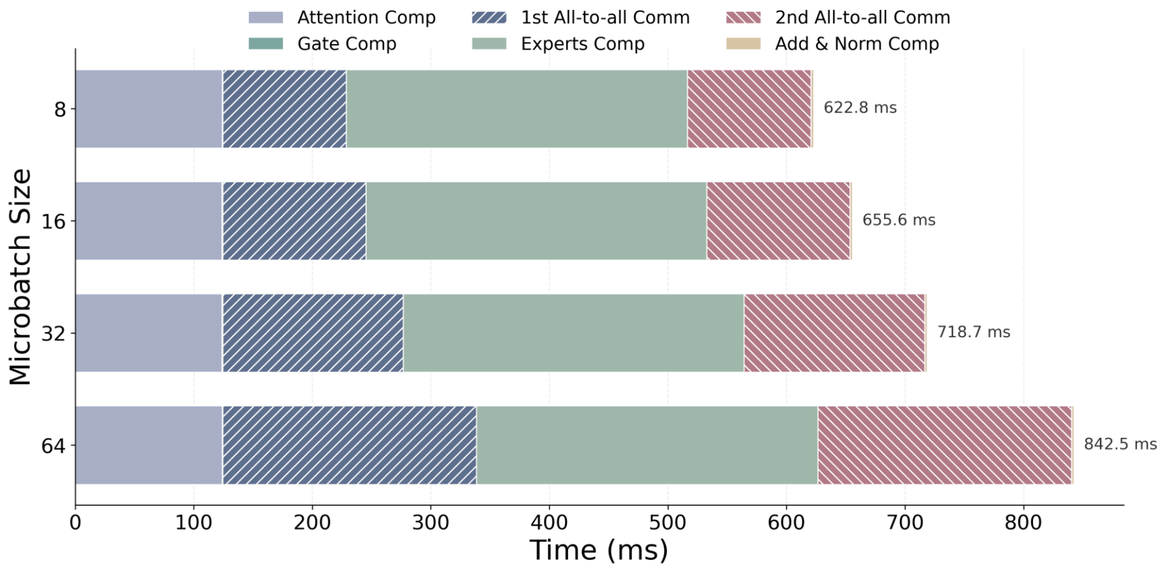
Attention

ht-sim
network simulator

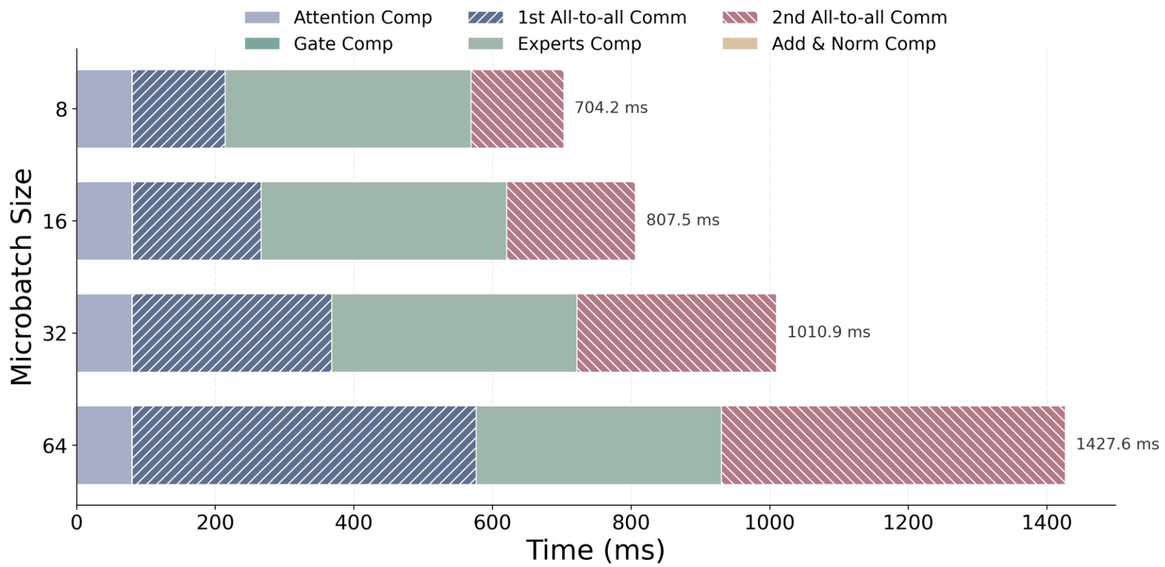
How long does it take for each kernels?

How does all-to-all communication occupy in the total iteration time?

Profile LLM MoE models



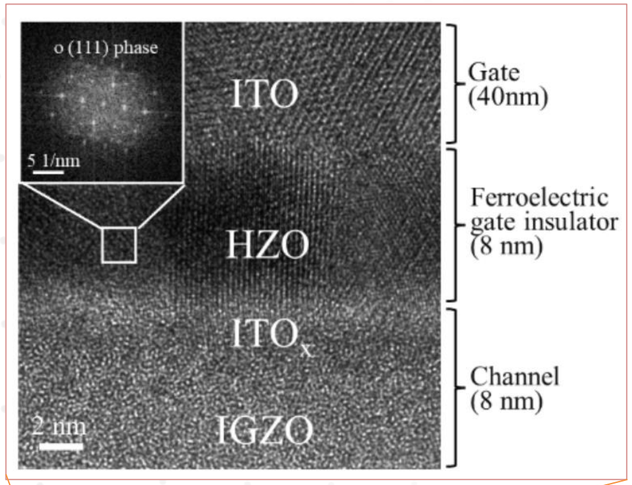
Mixtral 8x7B MoE Model single layer forward pass profiling in 128 Nvidia A100 GPUs with 400Gbps network



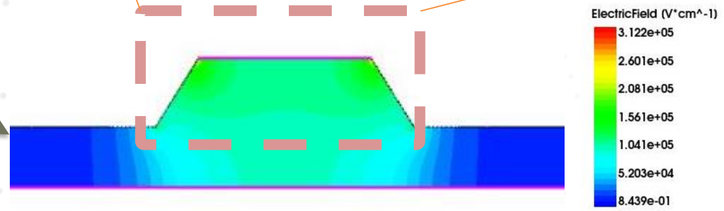
LLaMA-MoE 6.7B Model single layer forward pass profiling in 64 Nvidia A100 GPUs with 400Gbps network

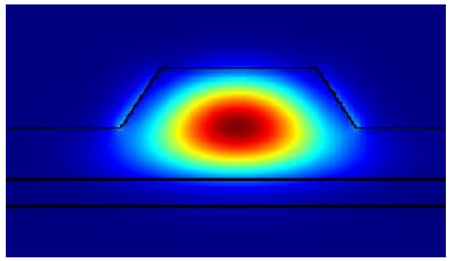
Expert Parallelism’s All-to-All communications occupy 1/3 to 1/2 of the total training iteration time

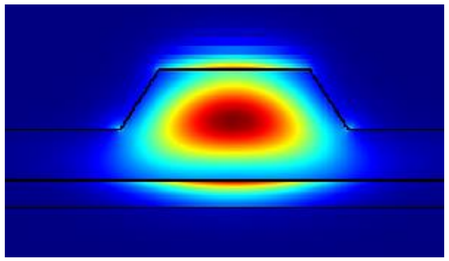
Ferroelectric MRR device simulation



Lithium Niobate Waveguide







TCAD

Electric field

Polarization

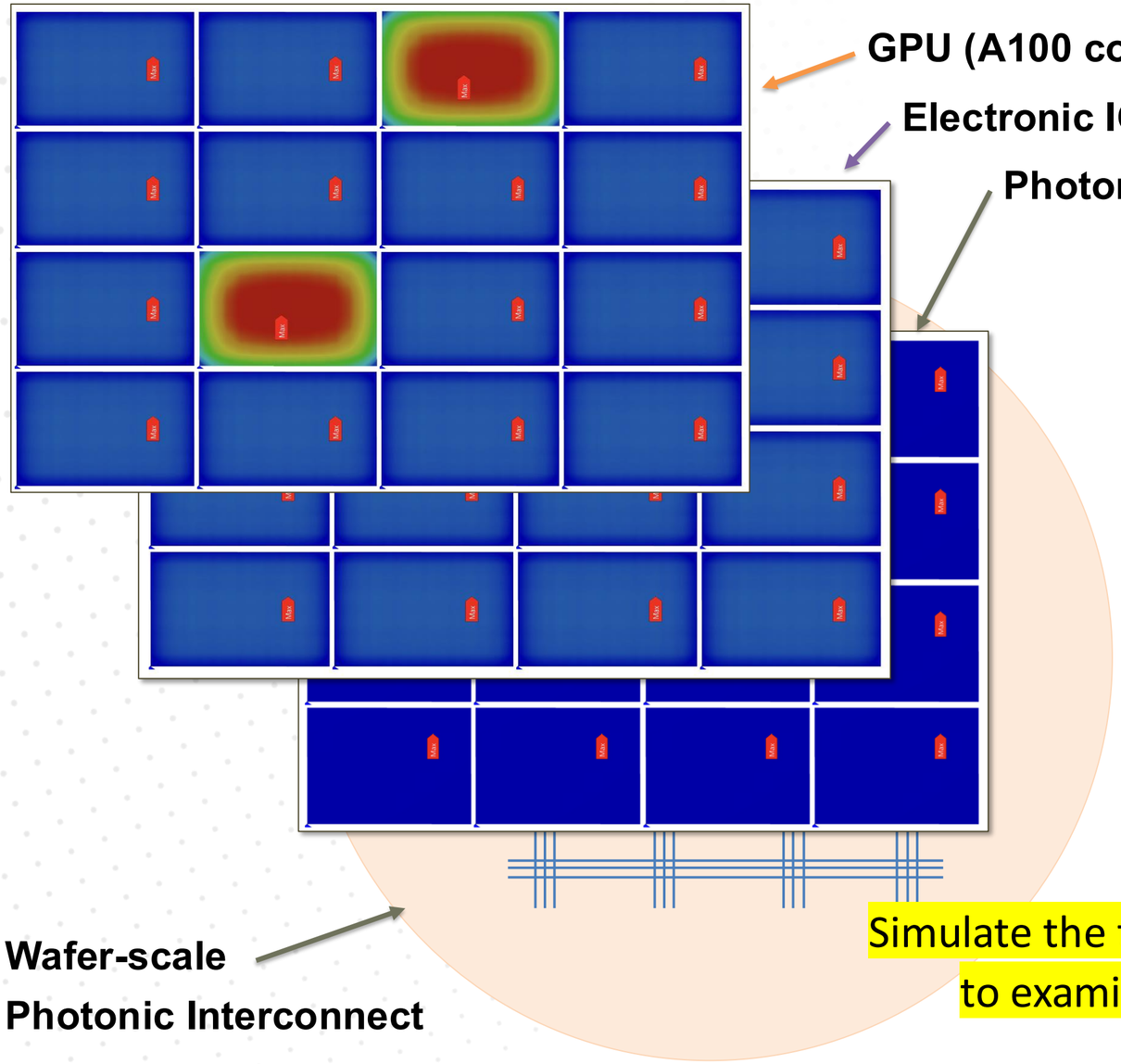
refractive index change

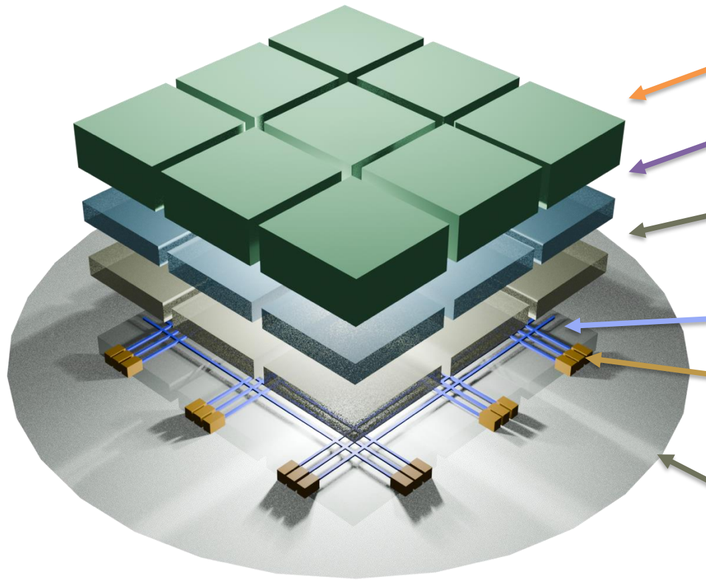
Ansys Lumerical

TE mode

TM mode

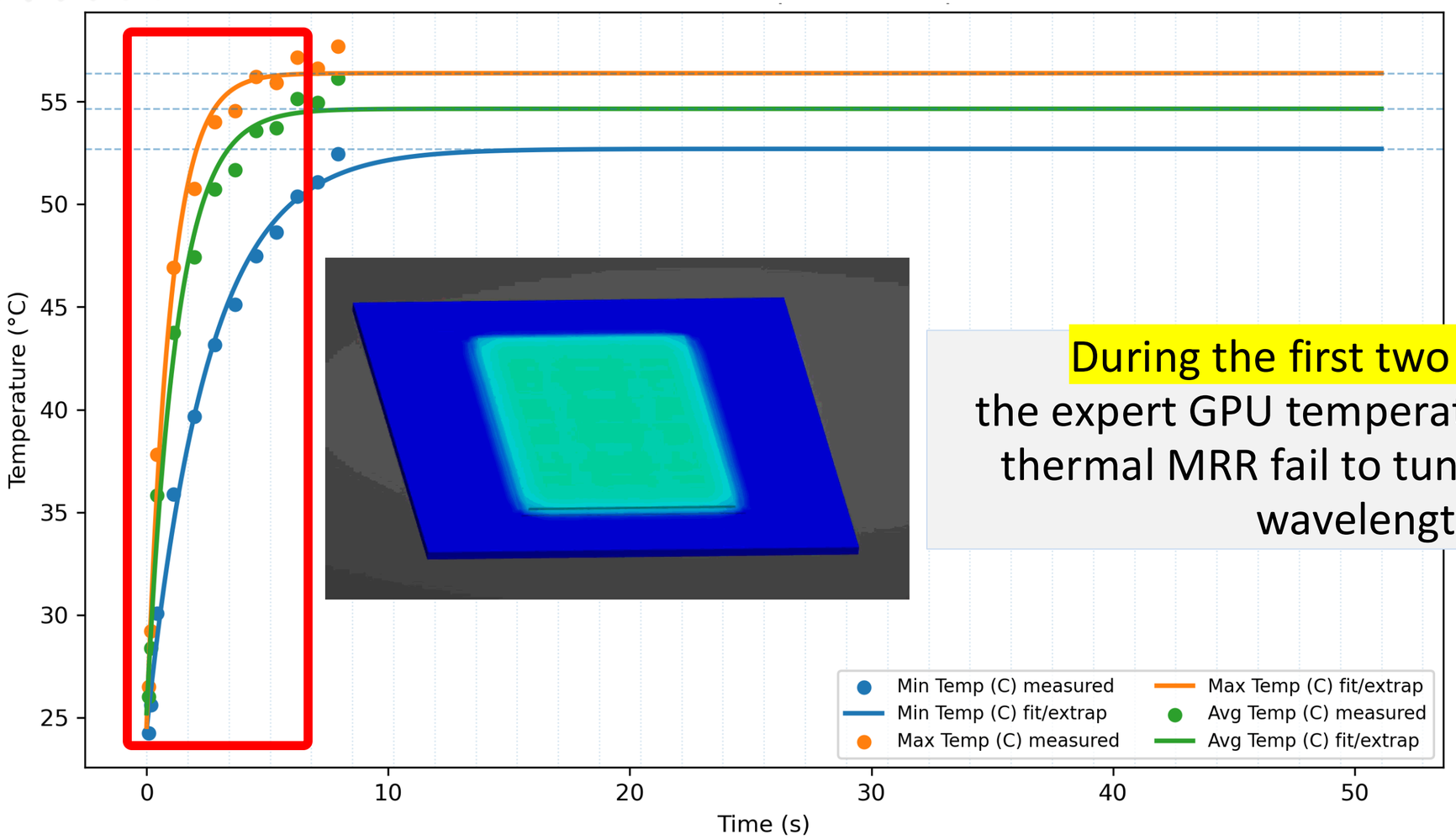
Thermal transient simulation with Ansys Thermal





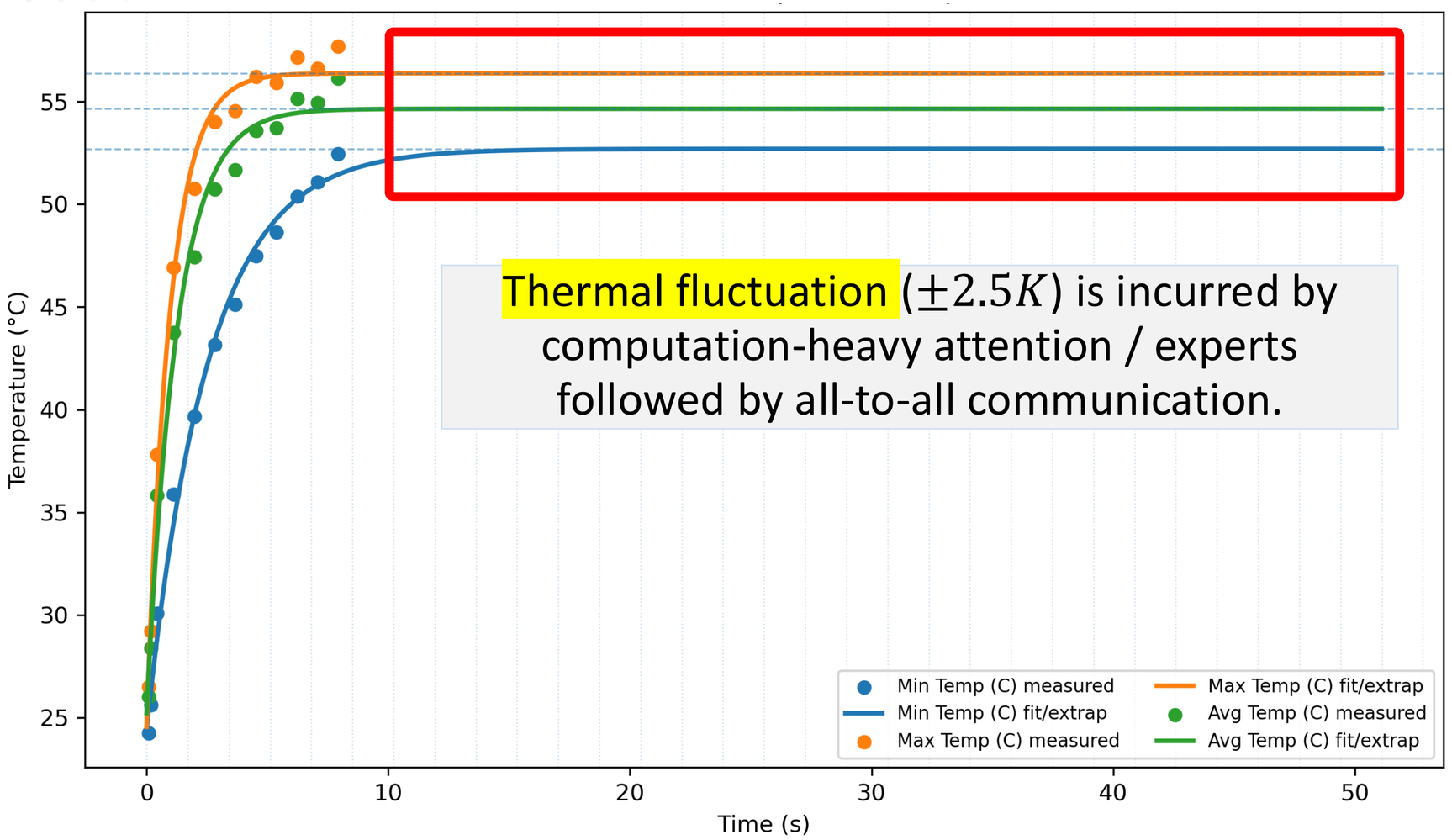
Simulate the thermal crosstalk between top GPU and bottom PIC to examine the temperature of MRR in the photonic IC.

Thermal transient response in expert GPU - MRR



During the first two MoE block,
the expert GPU temperature quickly rise,
thermal MRR fail to tune to resonance wavelength.

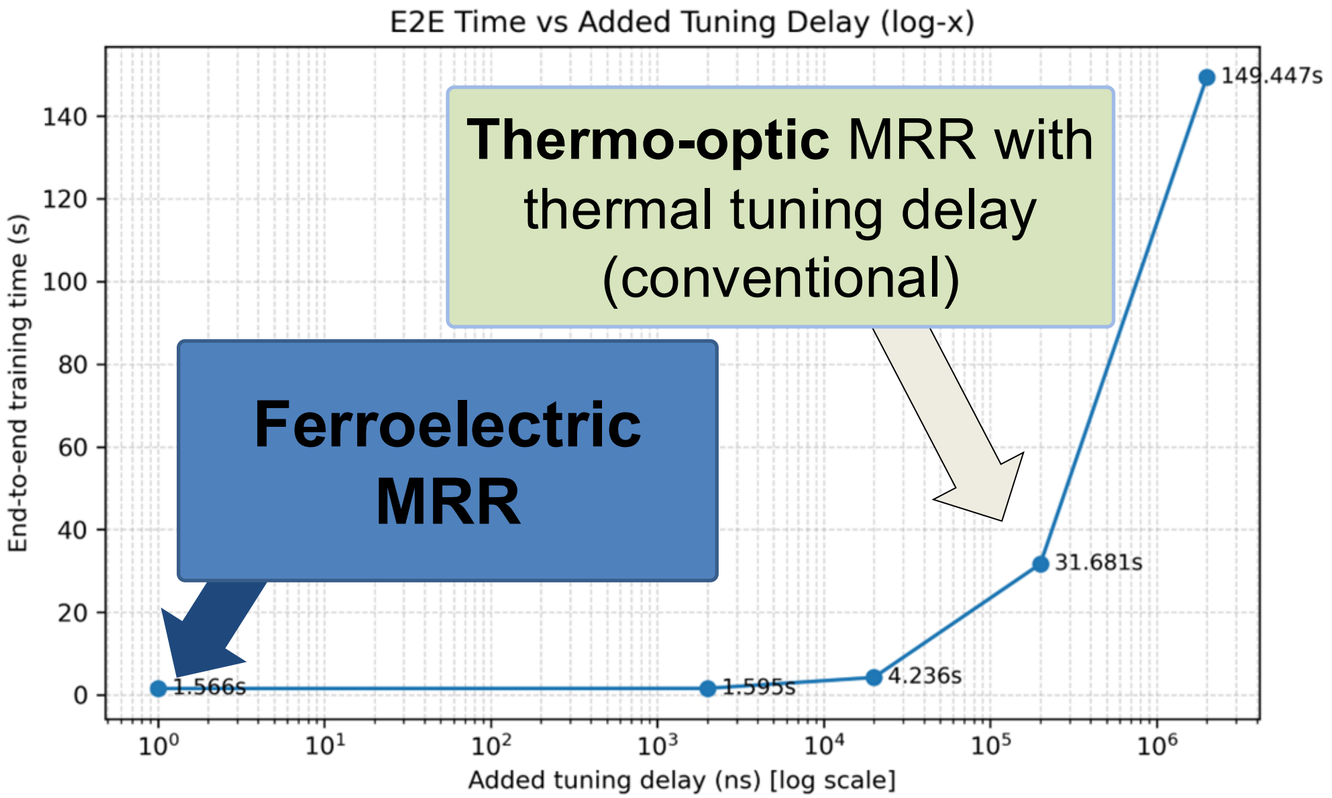
Thermal transient response in expert GPU - MRR



Thermal fluctuation ($\pm 2.5K$) is incurred by computation-heavy attention / experts followed by all-to-all communication.

Therefore, thermal MRRs (require $\pm 0.1K$) takes longer all-to-all communication time as it require relaxation time for GPU to reach steady-state and tune operating wavelength.

Training speed-ups for LLM MoE model



Sharp decrease in end-to-end training iteration time in LLM MoE model with ferroelectric MRRs.

Recap

WE TALKED ABOUT ..

GenAI LLM Era

LLM / Mixture-of-Experts
Distributed Training
All-to-All traffic pattern

Distributed Infrastructure for ML

Datacenter Infrastructure
Communication Hierarchy

Optical Interconnect Solutions

Optical Device Module

1. Co-Packaged Optics (CPO)
2. Optical Interposer

Ferroelectric MRR in Wafer/Panel-Scale Optical Interconnect Platform

Reference

- ECE 8803 HML Lecture Note by Prof. Divya Mahajan in Georgia Institute of Technology

- HOTCHIPS 2025 AI Rack Tutorial, Frank Helms (AMD) “How AI workloads shape rack system architecture”
- HOTCHIPS 2025 AI Rack Tutorial, Darrin Valis (AMD) “Scaling Fabric Technologies”
- HOTCHIPS 2025 AI Rack Tutorial, John Norton (Nvidia) “Case study : Nvidia GB200NVL72”
- HOTCHIPS 2025 AI Rack Tutorial, Gilad Shainer (Nvidia) “Co-Packaged Silicon Photonics Switches for Gigawatt AI Factories”
- HOTCHIPS 2023, 2025 AI Rack Tutorial, Lightmatter
- HOTCHIPS 2025 AI Rack Tutorial, Celestial AI

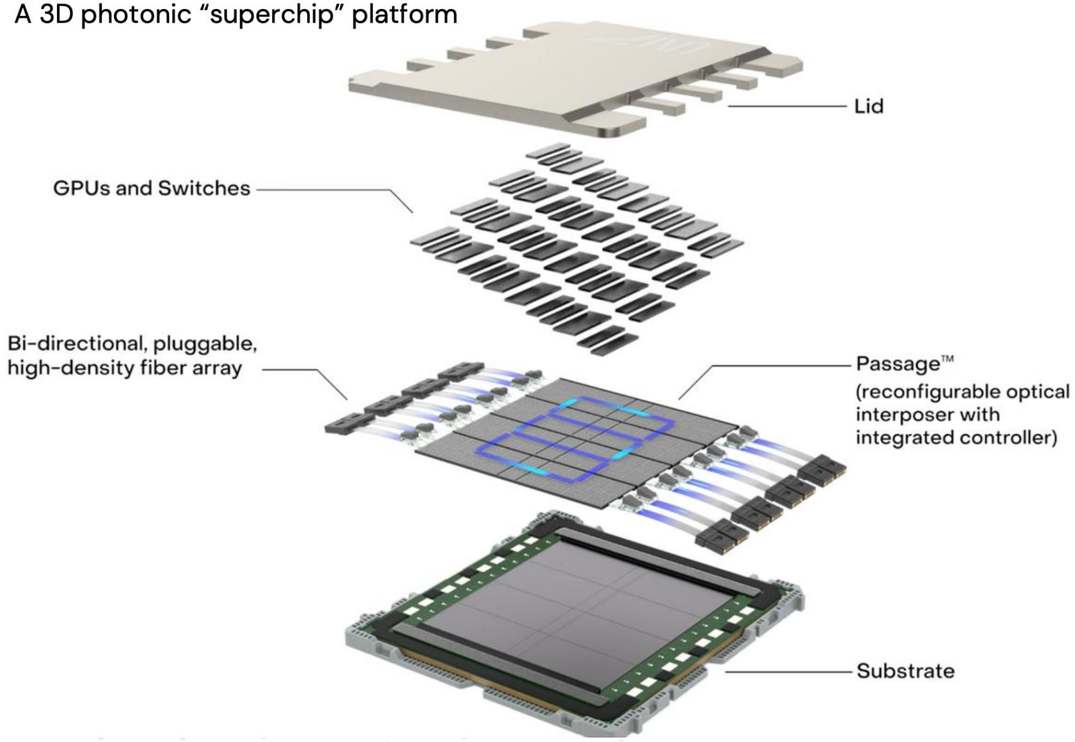
- Zhongkai Yu et al., ”Orders in Chaos: Enhancing Large-Scale MoE LLM Serving with Data Movement Forecasting”, *arXiv (2025)*
- Xudong Liao et al., “Mixnet: A Runtime Reconfigurable Optical-Electrical Fabric for Distributed Mixture-of-Experts Training”, ACM/IEEE SIGCOMM (2025)
- John H. Lau et al., “Co-Packaged Optics—Heterogeneous Integration of Photonic Integrated Circuits and Electronic Integrated Circuits” (2025)

Thank you

Additional Slides

Passage M1000

A 3D photonic “superchip” platform



Lightmatter also manufacture wafer scale optical interconnect substrate