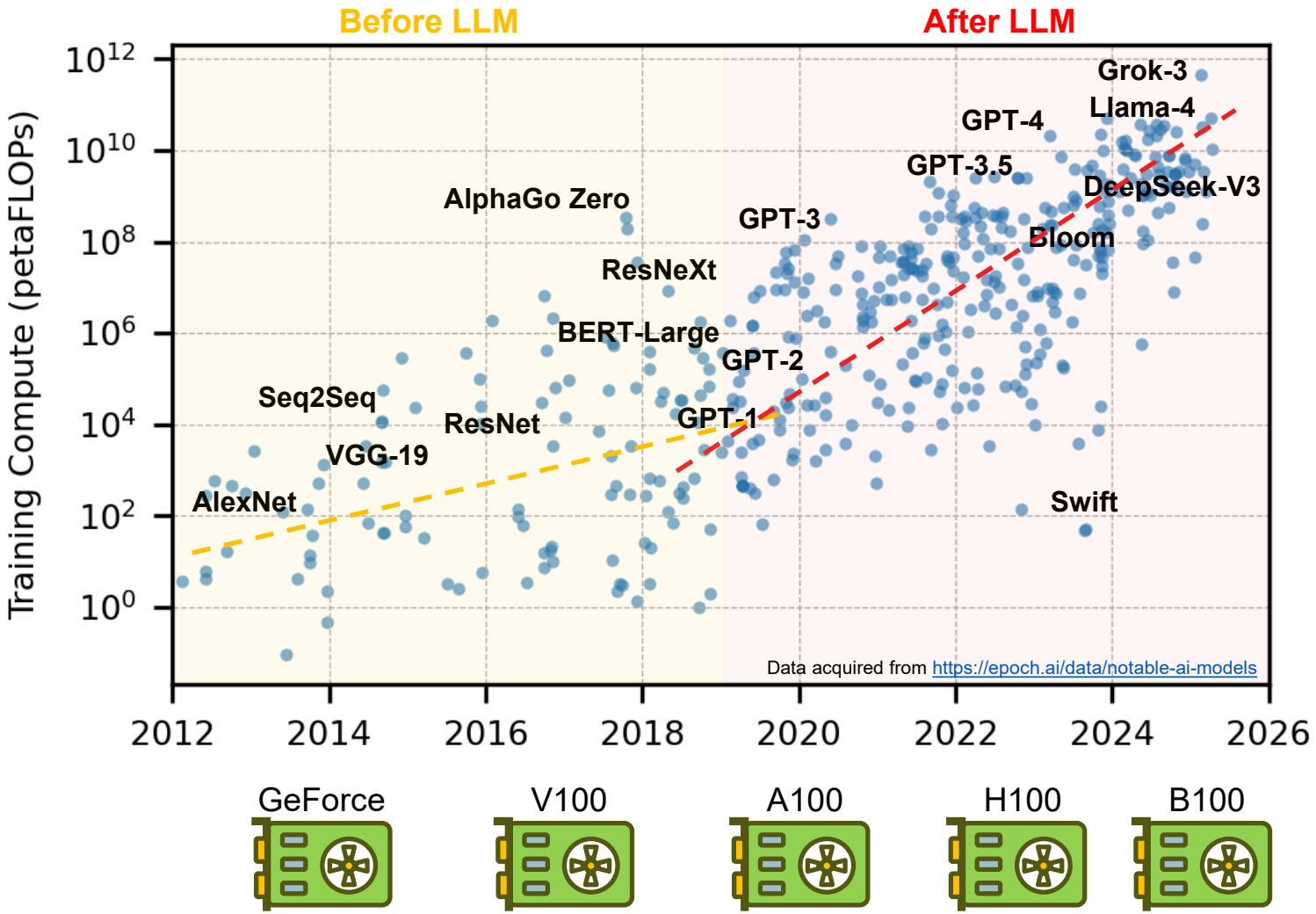
Photonic Connectivity for AI Systems

Keren Bergman

Columbia University,
New York, NY 10027, USA

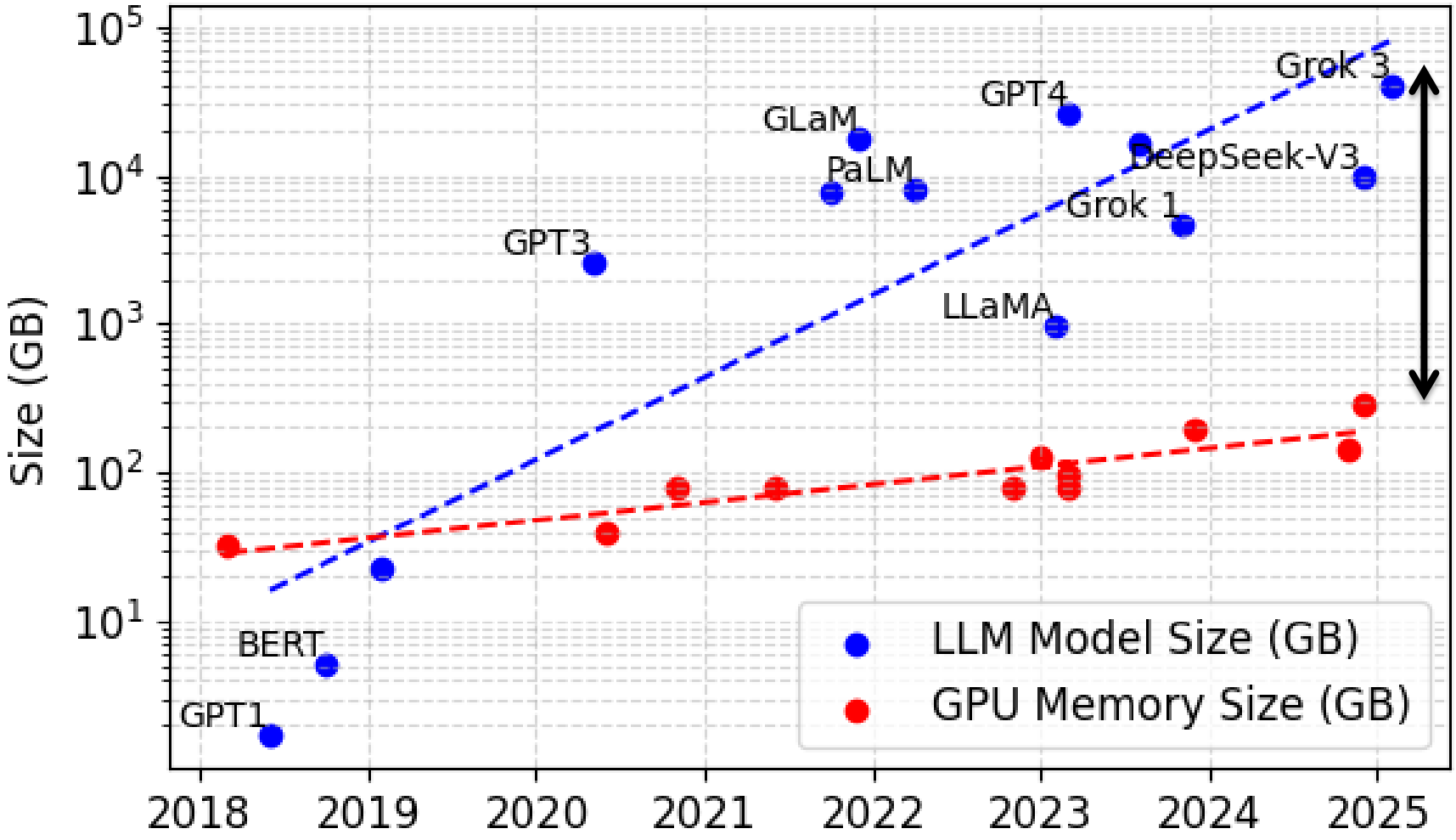
Artificial Intelligence Model Scaling Trend



- **Inflection Point (2018–2020)**
  - Shift from conventional deep learning (DL) to transformer-based (LLM)
- **Steady increase in compute**
- **Co-Scaling Compute and Communication**

https://developer.nvidia.com/blog/transform-the-data-center-for-the-ai-era-with-nvidia-dpus-and-nvidia-doca/

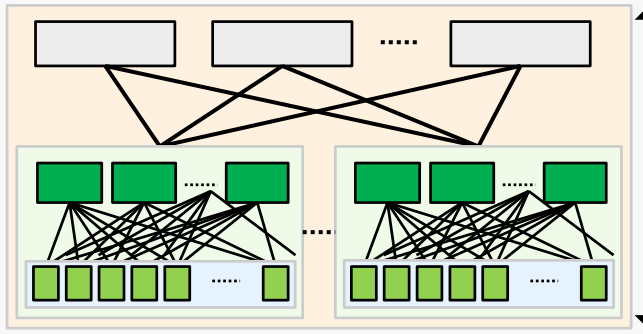
Growing Gap: Model Sizes Exceed GPU Capacity



> 2 orders of mag

Network Hierarchies in Accelerator Based AI Compute Systems

Multi-Stage Hierarchical Network

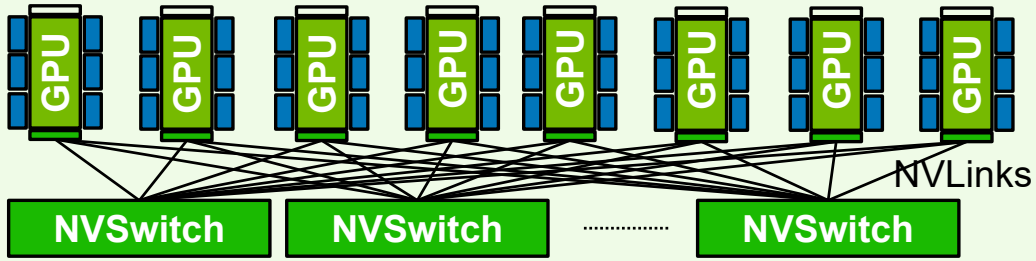


80× Bandwidth Taper across system

- Communication bottleneck across hierarchies
- Increased complexity in model deployment

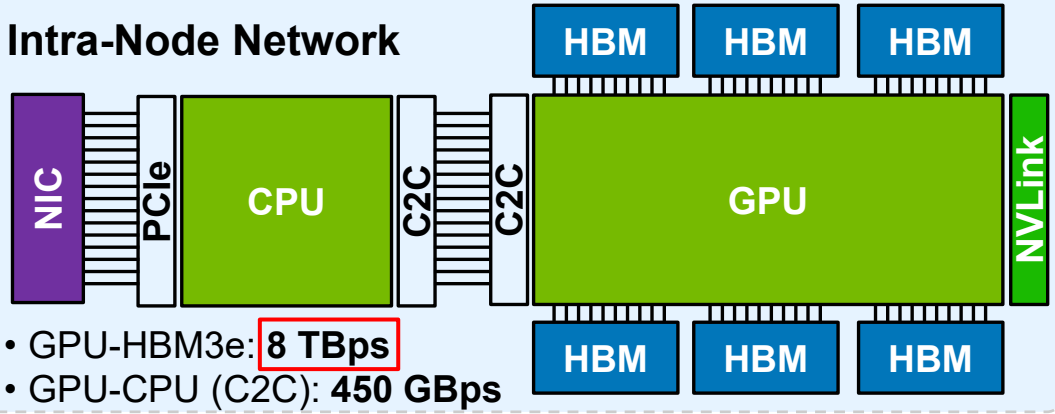
Inter-Node (Scale-Up) Network

- NVLink 5.0: **900 GBps** / GPU
- Up to 256 GPUs under same NVLink domain



Photonic Interconnect Design Space

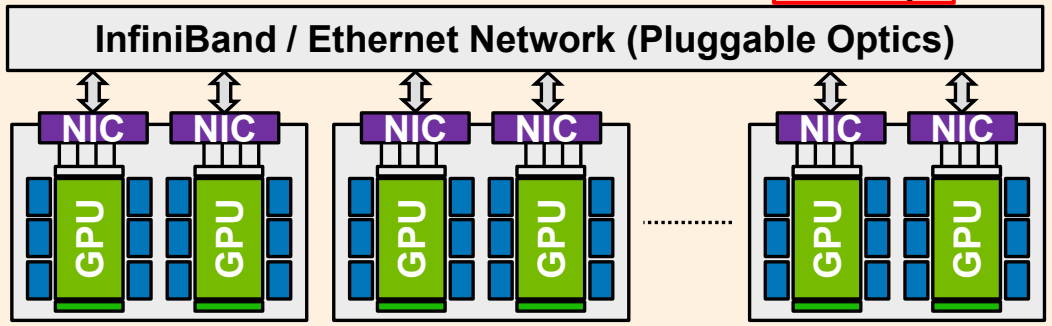
Intra-Node Network



- GPU-HBM3e: **8 TBps**
- GPU-CPU (C2C): **450 GBps**

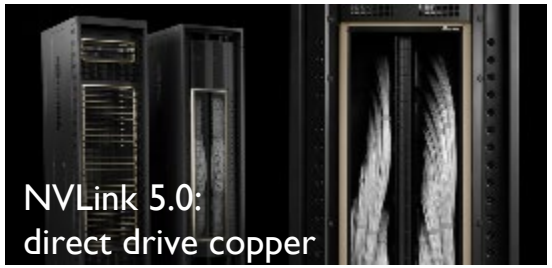
Inter-Node (Scale-Out) Network

- NIC: **800 Gbps** / GPU



Commercial Photonics in Scale Up

Nvidia's GB200 NVL72



- Copper-based electrical links
- Limited scale-up domain (72)
- Nvidia's higher per-chip compute is constrained by its electrical interconnects when scaling for aggregate compute performance.

Huawei CloudMatrix 384

- Linear Pluggable Optical cables
- Increased scale-up domain (384)
- Higher total compute power due to larger scale-up domain (with lower per-GPU compute)
- Lower power-per-bit

Scaling AI Networking Infrastructure

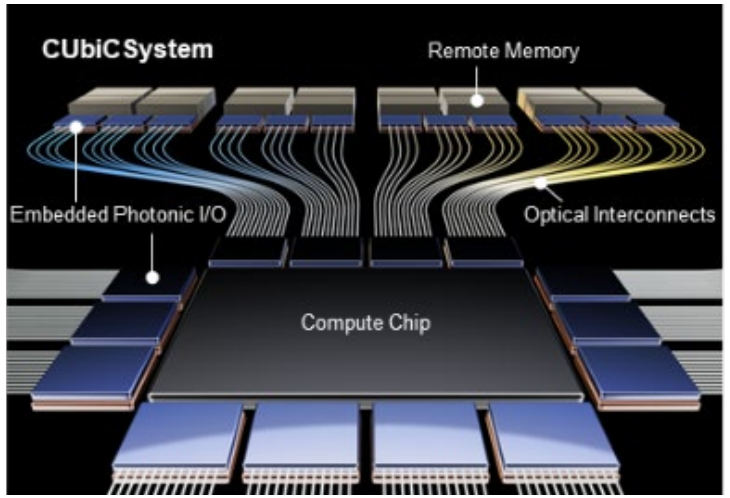
Compute / Memory / Interconnect Comparison*

Chip-Level	Nvidia GB200	Ascend 910C
TFLOPs	2,500	780
HBM Capacity (GB)	192	128
HBM Bandwidth (TBps)	8	3.2
Scale Up Bandwidth (Tbps)	7.2	2.8
Scale Out Bandwidth (Tbps)	0.4	0.4

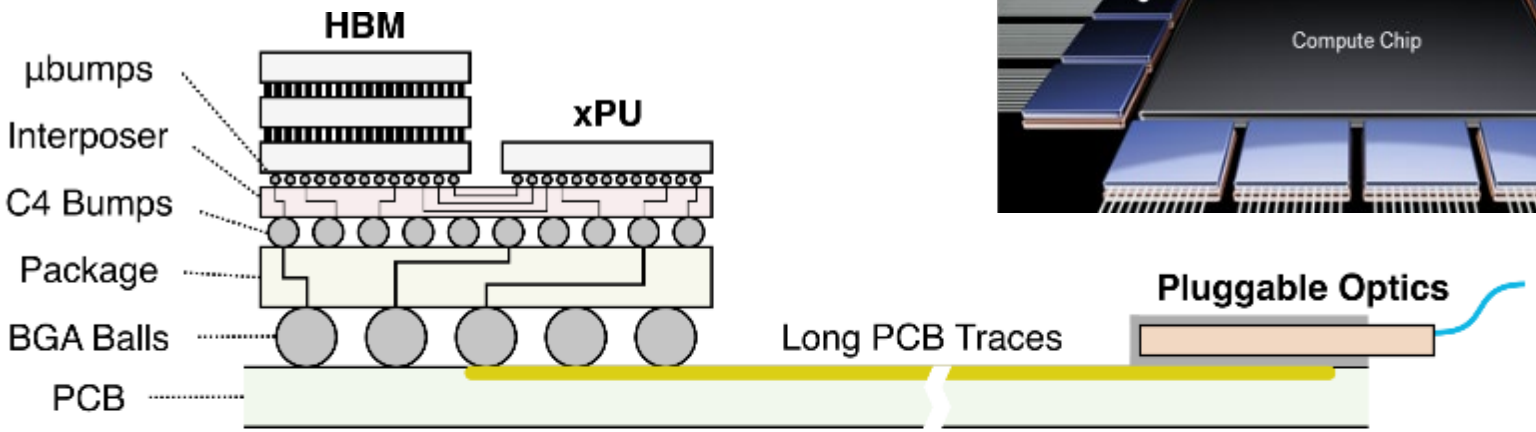
System-Level	GB200 NVL72	CLoudMatrix 384
# Compute Units	72	384
PFLOPs	180	300
All-In System Power (kW)	145	599
HBM Capacity (TB)	13.8	49.2
HBM Bandwidth (TBps)	576	1,229
Scale Up Bandwidth (TBps)	64.8	134.4
Scale Out Bandwidth (TBps)	3.6	19.2

* Numbers based on SemiAnalysis report: https://semianalysis.com/2025/04/16/huawei-ai-cloudmatrix-384-chinas-answer-to-nvidia-gb200-nvl72/
** Estimated based on 384 compute-unit scale-up domain size

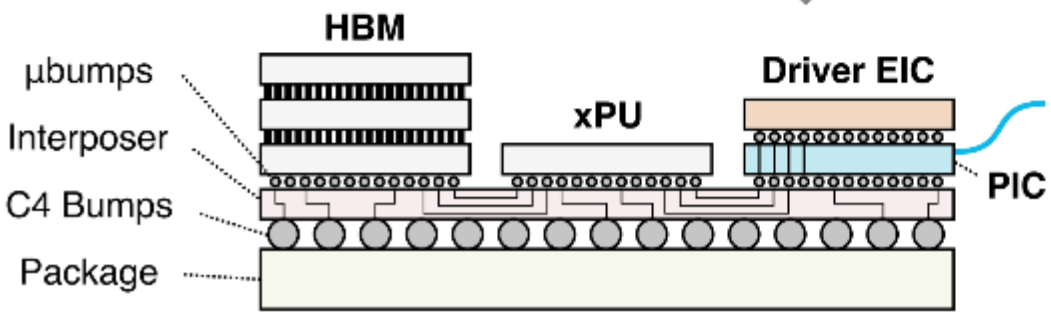
Bringing Photonics into Computing Sockets



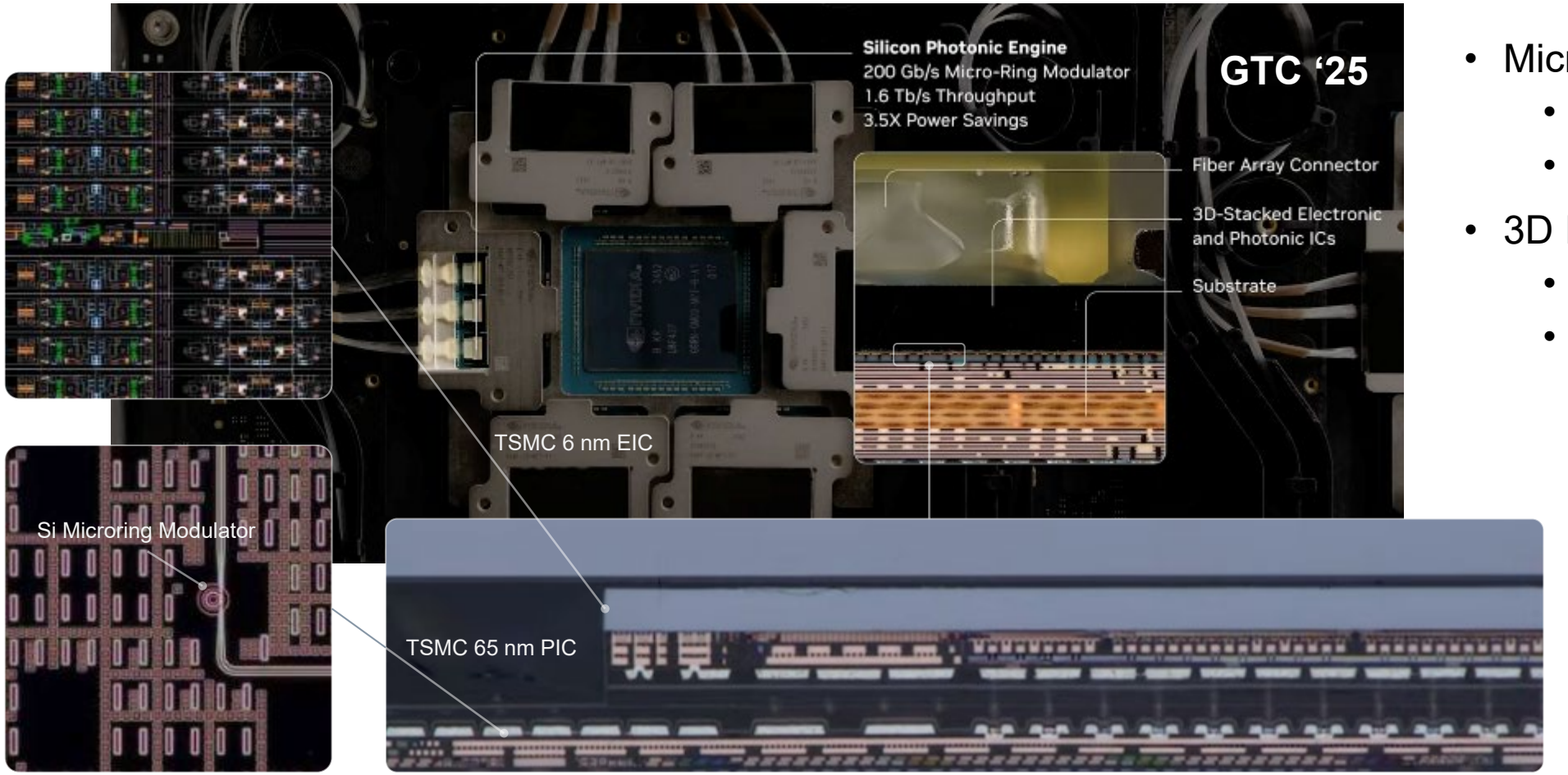
Today's Optical Connectivity



Embedded Photonics Data Input/Output (I/O)



NVIDIA's Co-Packaged Optics for Scale-Up



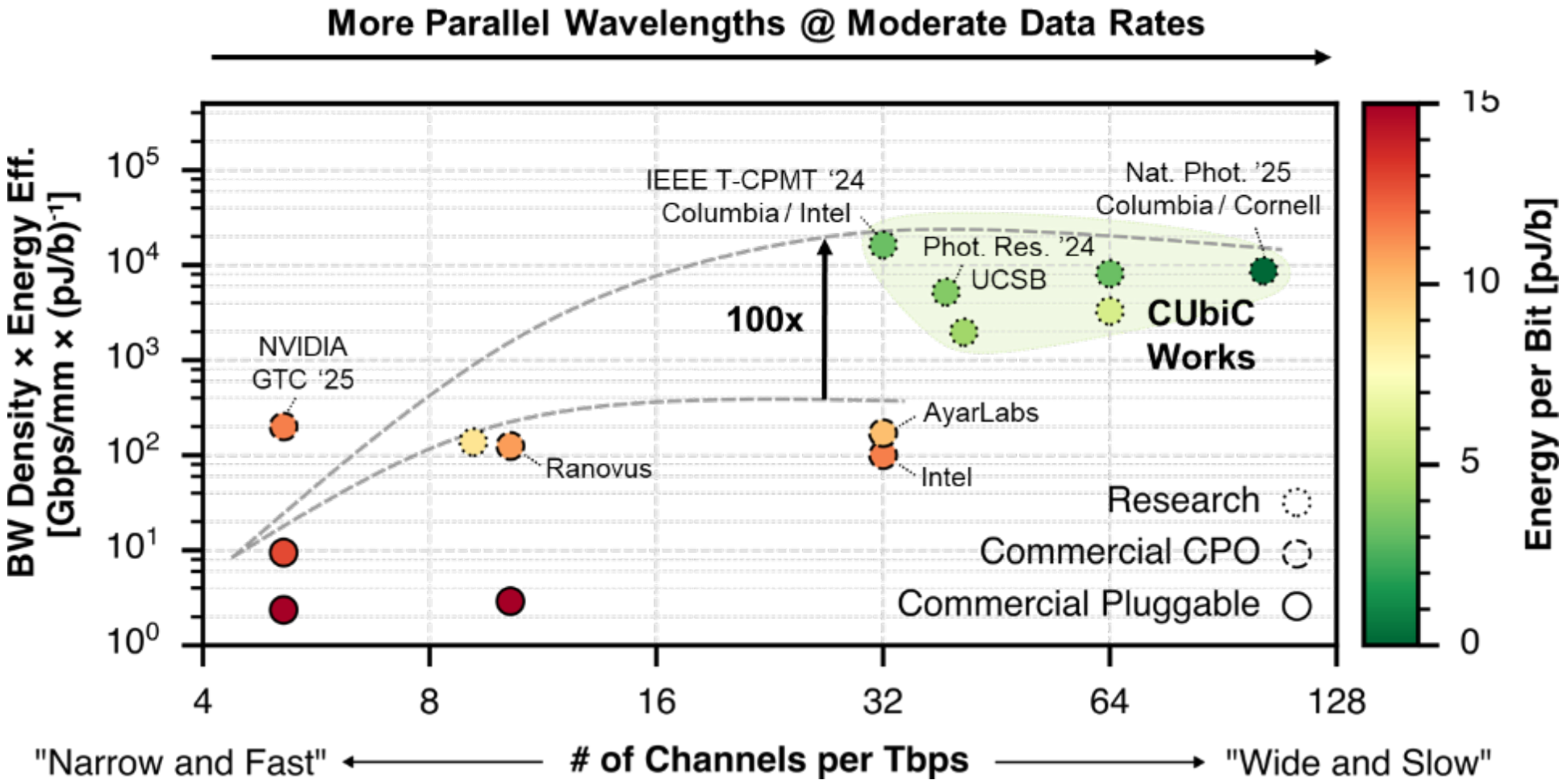
- Micro-resonators
  - Compactness
  - Scalability
- 3D Integration
  - Best of both worlds
  - Integration density

<u>What's still needed:</u>
Multi-λ Sources
Scalable DWDM Link

Validates DWDM Link Architecture for BW density

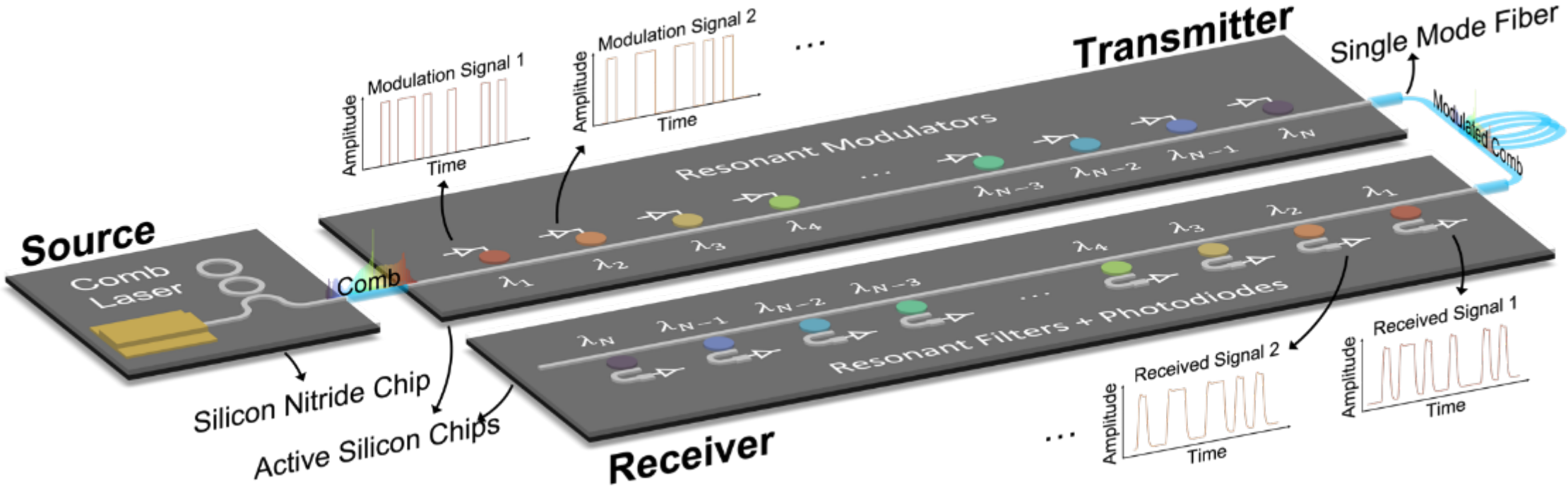
NVIDIA Silicon Photonics: https://www.nvidia.com/en-us/networking/products/silicon-photonics/

Realizing Extreme Bandwidth Density with Energy Efficiency



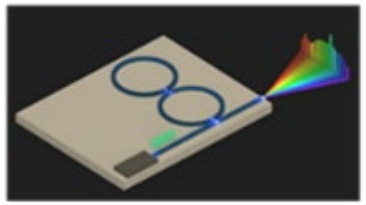
Link with Massive Parallelism in Wavelength Domain

- Realize extreme bandwidth density: multi-Tbps per single link
- Ultra – low energy/bit <1pJ/b

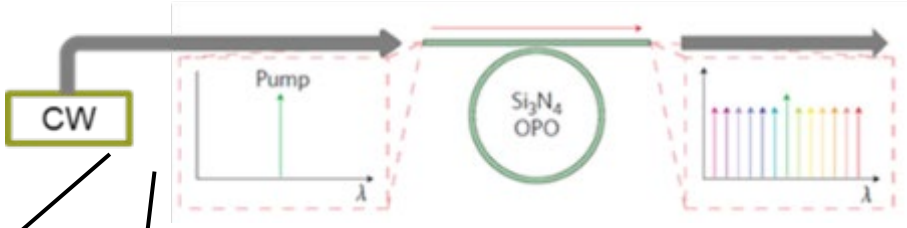


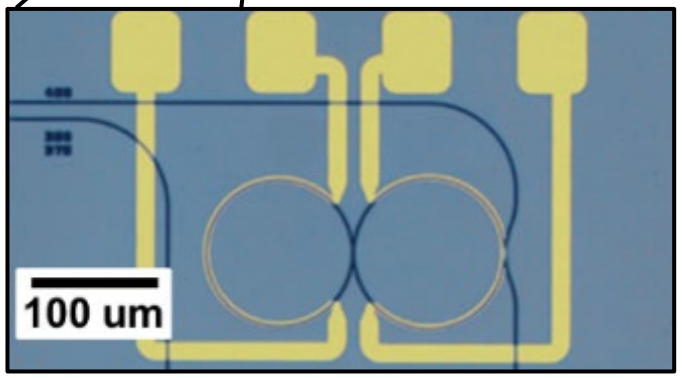
- Link bandwidth and energy/bit are distance independent

Key Enabler: High Power Comb

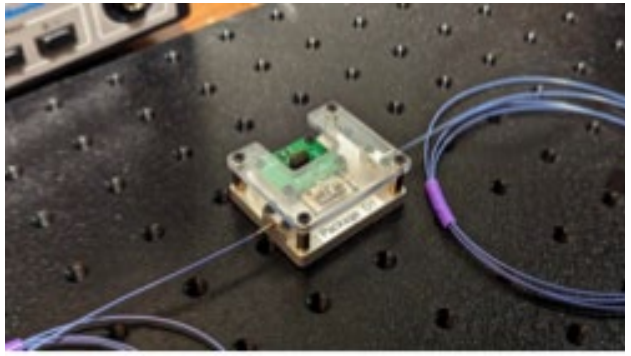


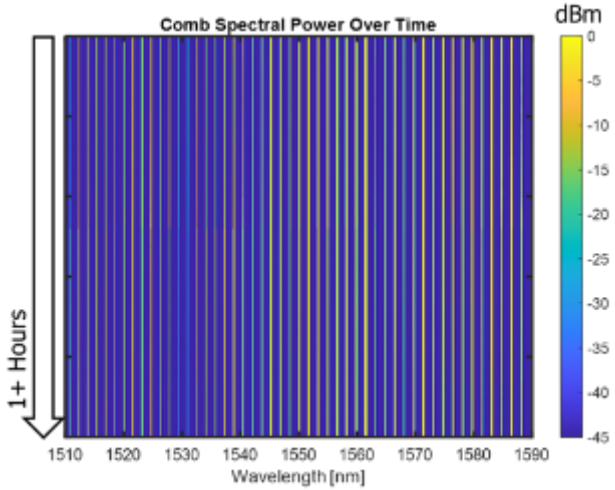
Frequency comb



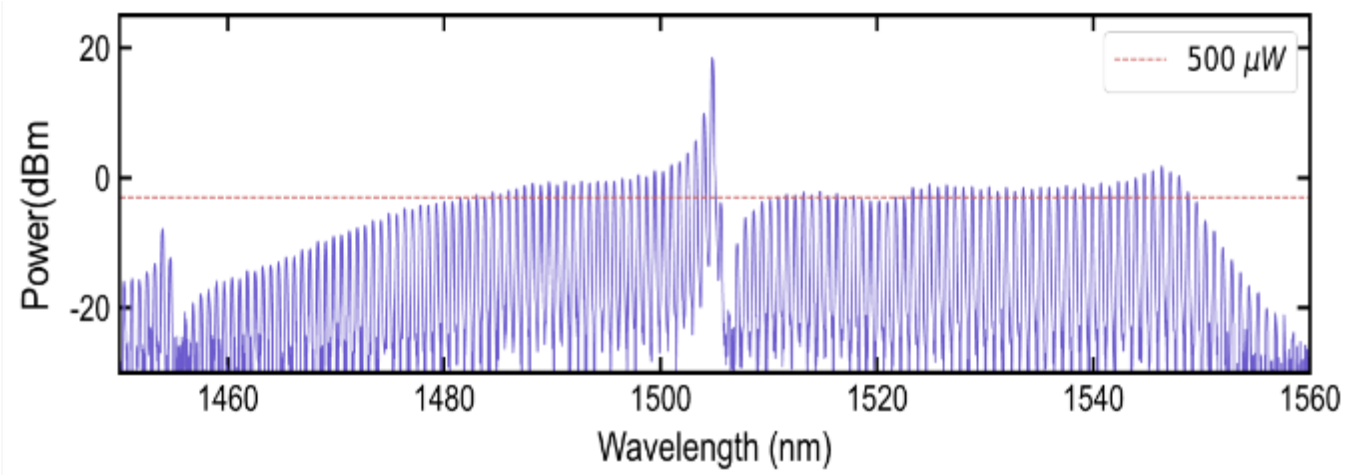


Gaeta, Lipson





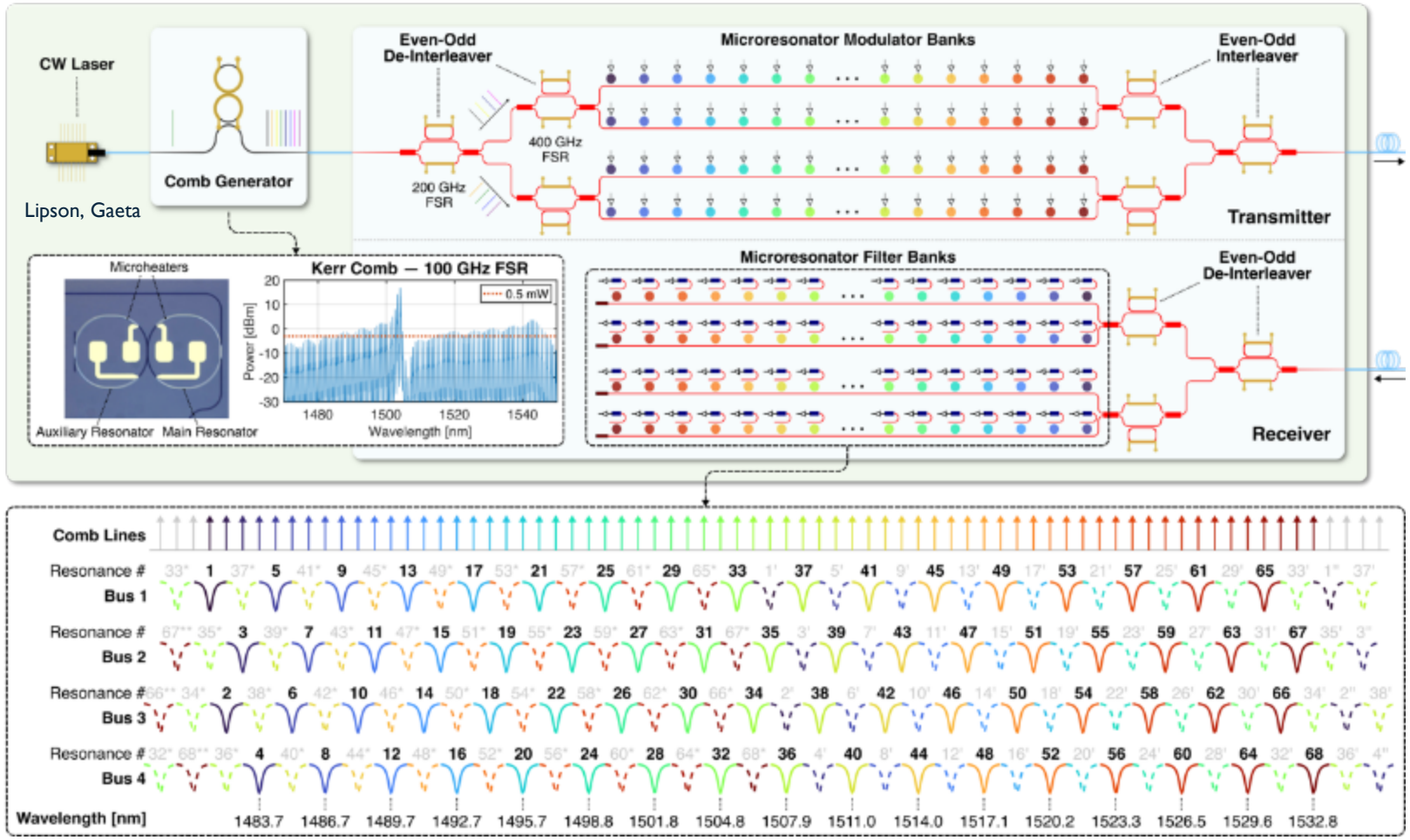
75 Channels >0.5mW



- Normal GVD Kerr comb with 46% conversion efficiency

Tbps/Fiber Link Co-Designed with Comb

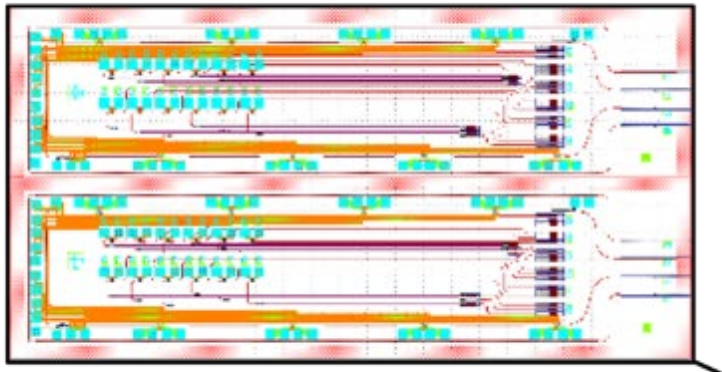
- 100 GHz comb source
  - 75 lines > 0.5 mW
  - 46.2% conversion efficiency
- Broadband interleavers
  - 2 stages 200/400 GHz FSR
  - Enables multi-FSR
- 16 Gbps/ch × 64 ⇒ 1 Tbps/link
- 32 Gbps/ch × 64 ⇒ 2 Tbps/link
- **32 Gbps/ch × 128 ⇒ 4 Tbps/link**



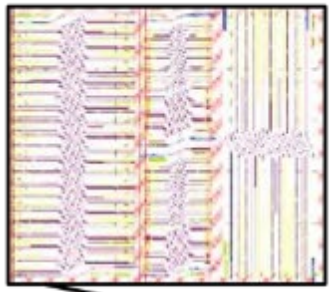
Yuyang Wang *et al. SPIE OPTO*, 2023
Yuyang Wang *et al. IEEE CICC*, 2024
Yuyang Wang *et al. IEEE T-CPMT, 2025*

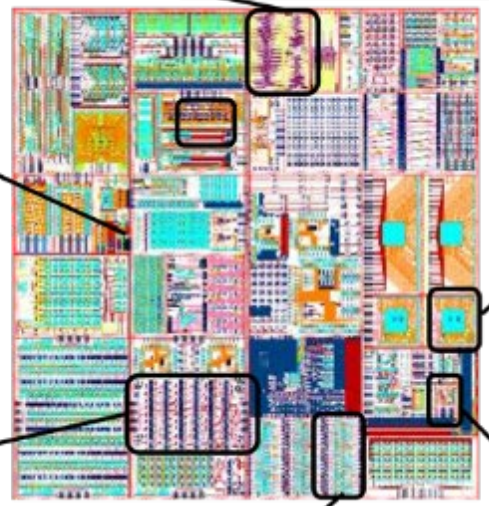
Full 300 mm Wafer - Cedar

FPGA-packaged WDM Transmitters

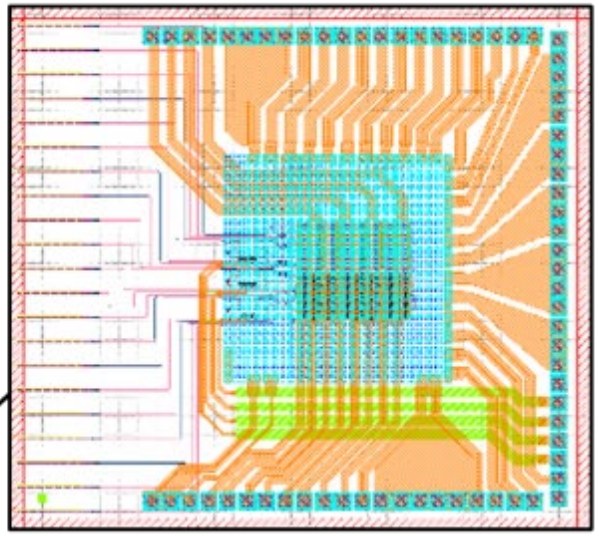


Sub-dB
Edge
Couplers

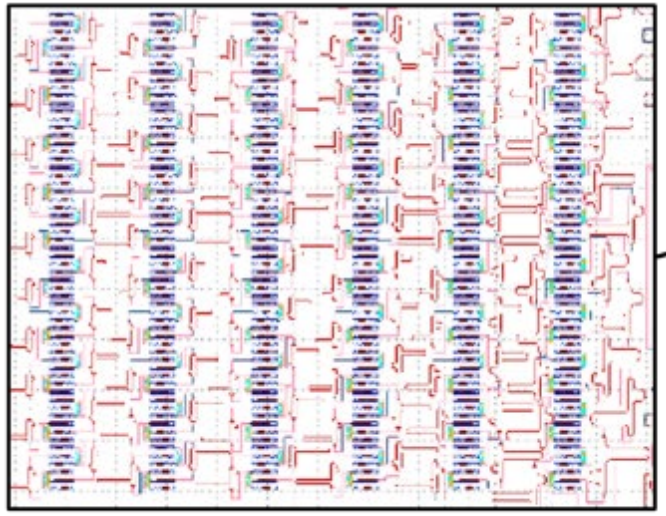




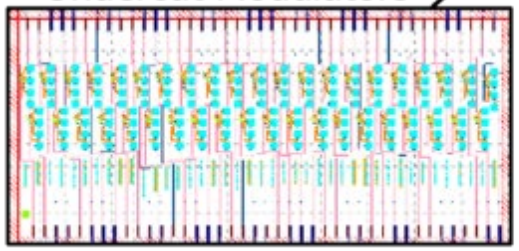
MCM with Custom Modulators



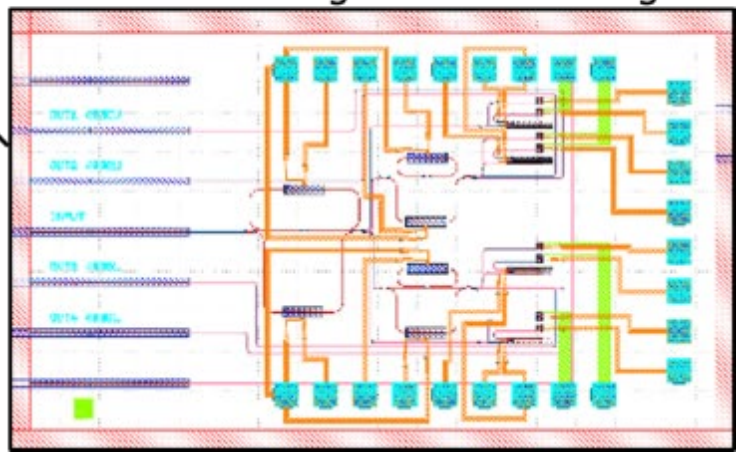
Wafer-scale Quantification of Fabrication
Robust Platform Phase Errors



Undercut Modulators

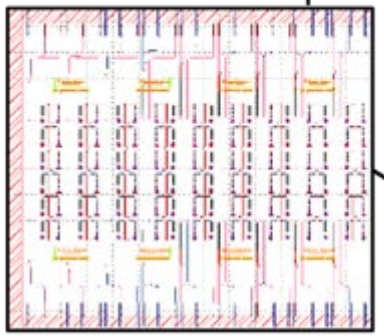


Cascaded RMZI Interleavers with
Automated Alignment & Tracking

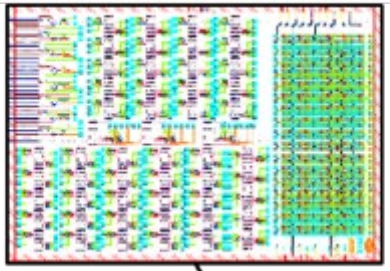


Full 300 mm Wafer - Oak

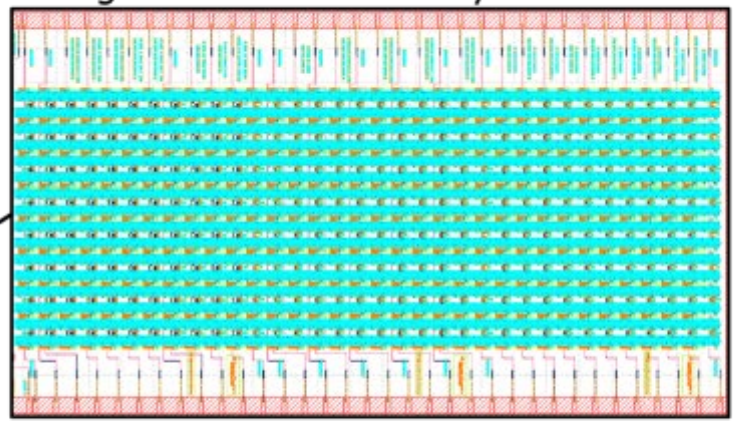
Polarization Rotator Splitters



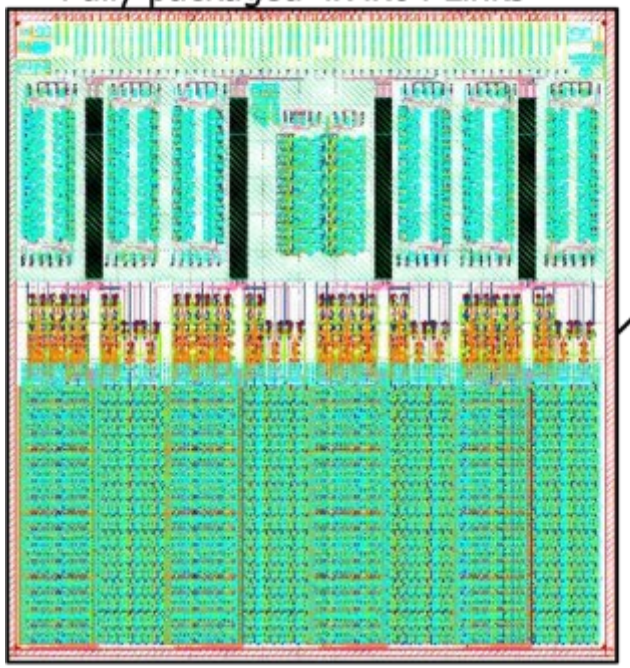
Non-packaged Link Test Structures

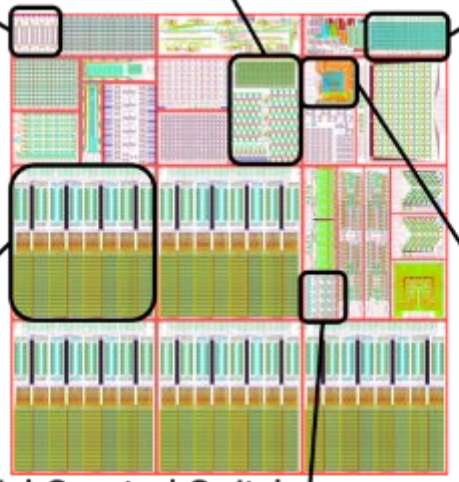


Package-able Resonator Arrays for Tx and Rx

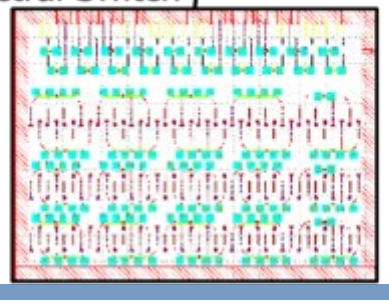


Fully packaged 4x4x64 Links

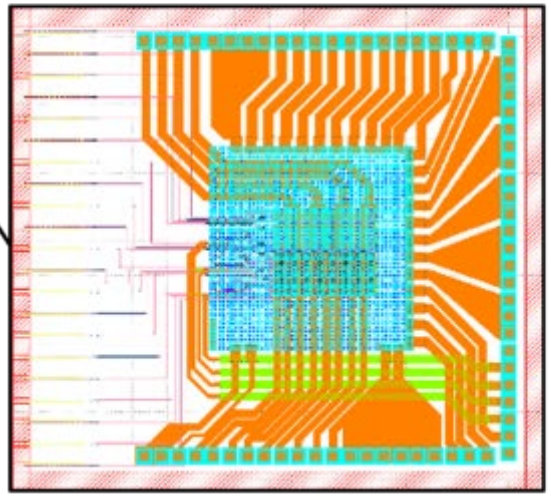




Novel Spatial-Spectral Switch Structures



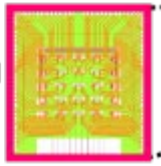
MCM with Updated ECs & Custom Resonators



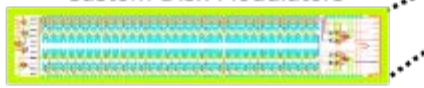
Full 300mm Wafer - Willow

Low Loss Photonic Computing for ML

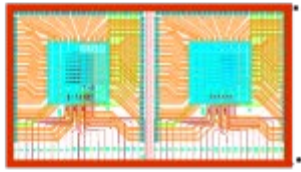
Pacakged Optical Switch



128 Channel TX with Custom Disk Modulators

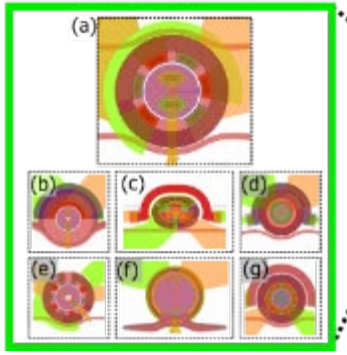


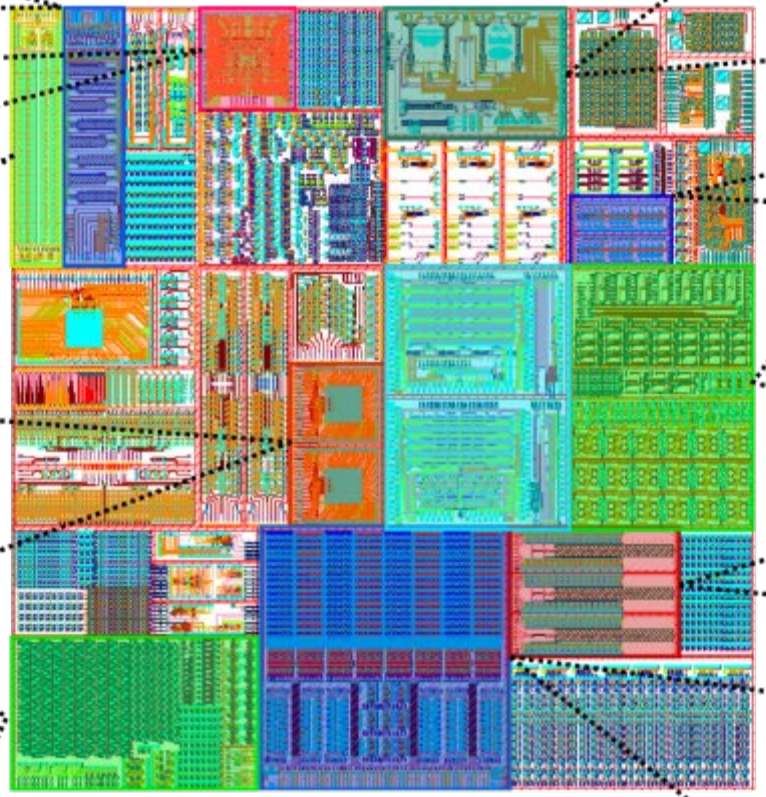
Updated MCM for 100 and 200 GHz Channel Spacing

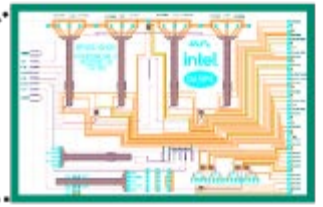


Custom Modulators:

a) Interleaved Junction Disk
b) Ridge Coupled Disk
c) THB Disk
d) Coaxial Junction Disk
e) Segmented PAM-4 Disk
f) Over coupled Ring Phase Shifter
g) Interleaved Disk with Half Etch Heater

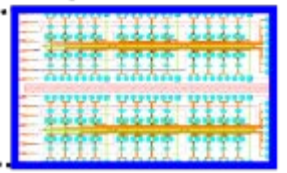


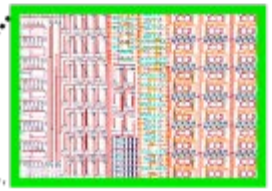




Reconfigurable Tranceiver for Space-Based Communications

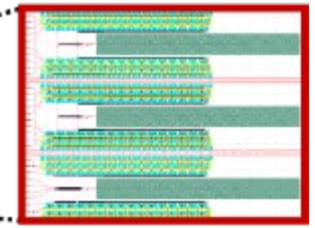
Ring Assisted MZMs

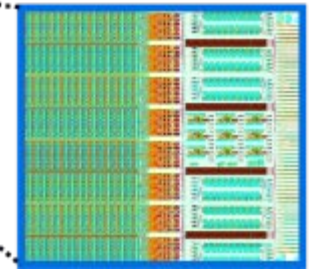




Ultra Low-Loss Photonic Escalators

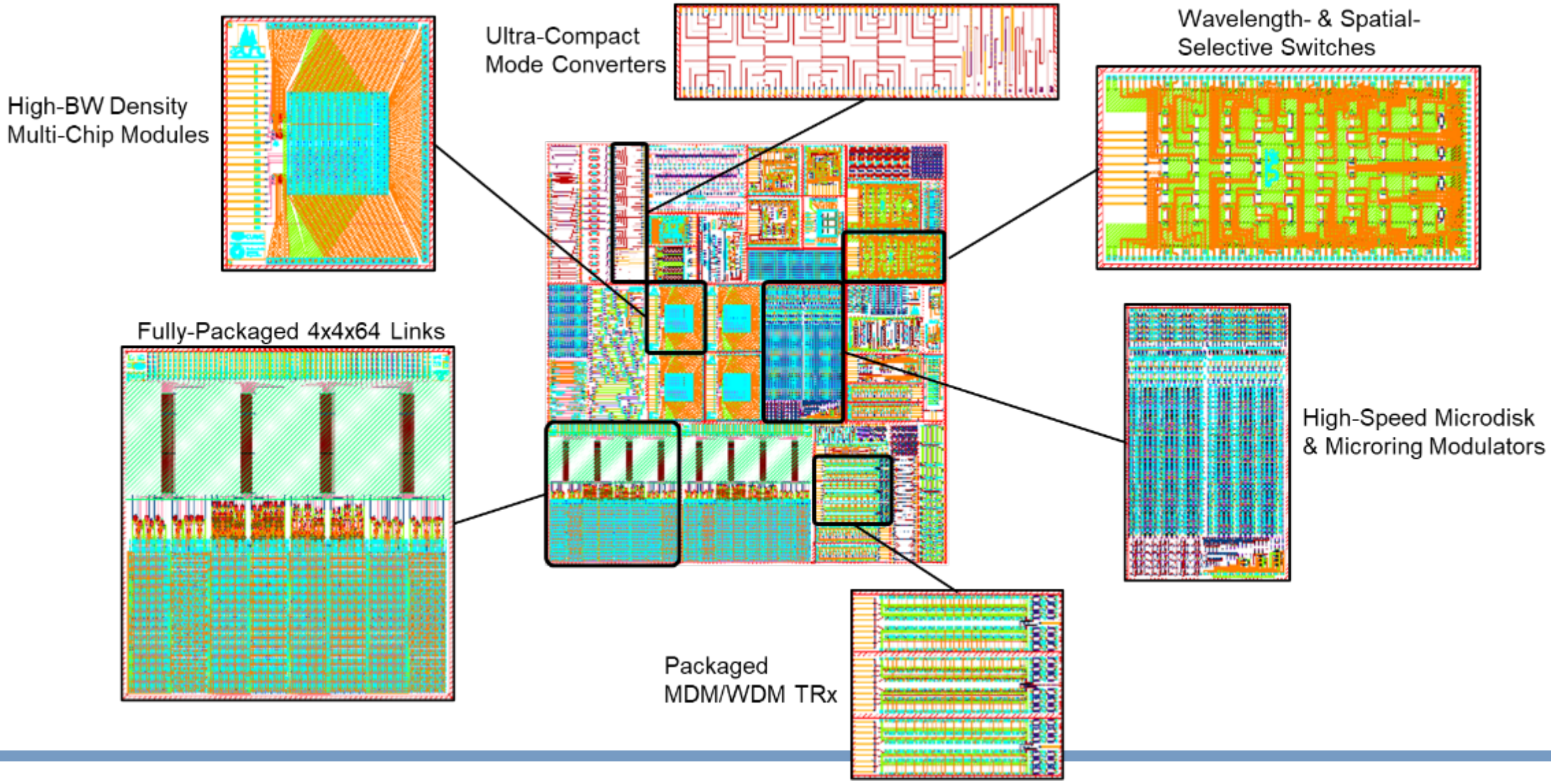
MDM + WDM Tranceivers





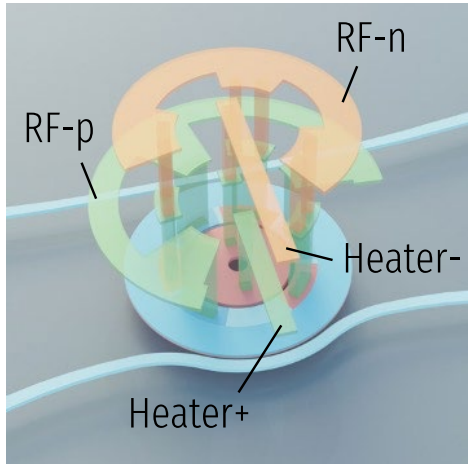
Intel PIPES Multi-Tbps Tranceiver

Full 300 mm Custom Wafer – Spruce (in Fab)

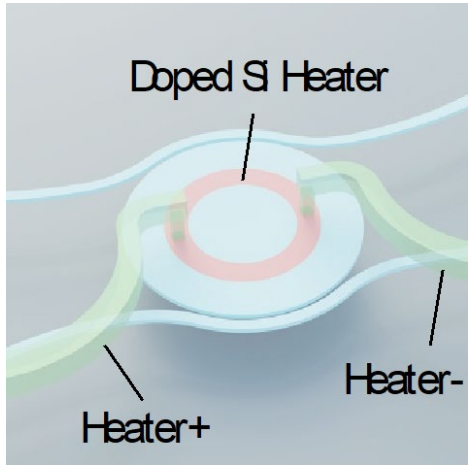


Link Multi-Channel WDM Design

Custom Resonators Co-Designed with Comb

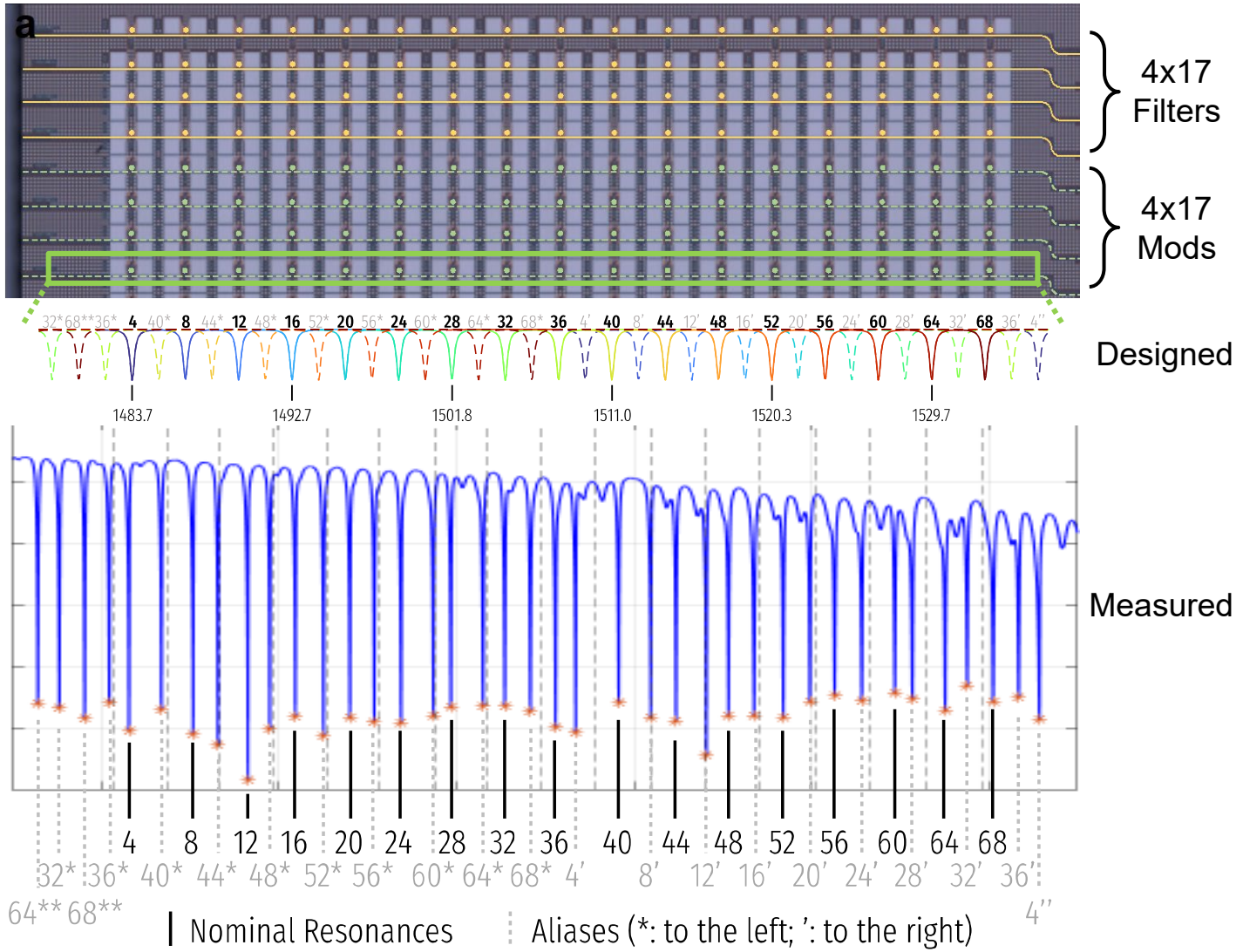


Microdisk Modulator

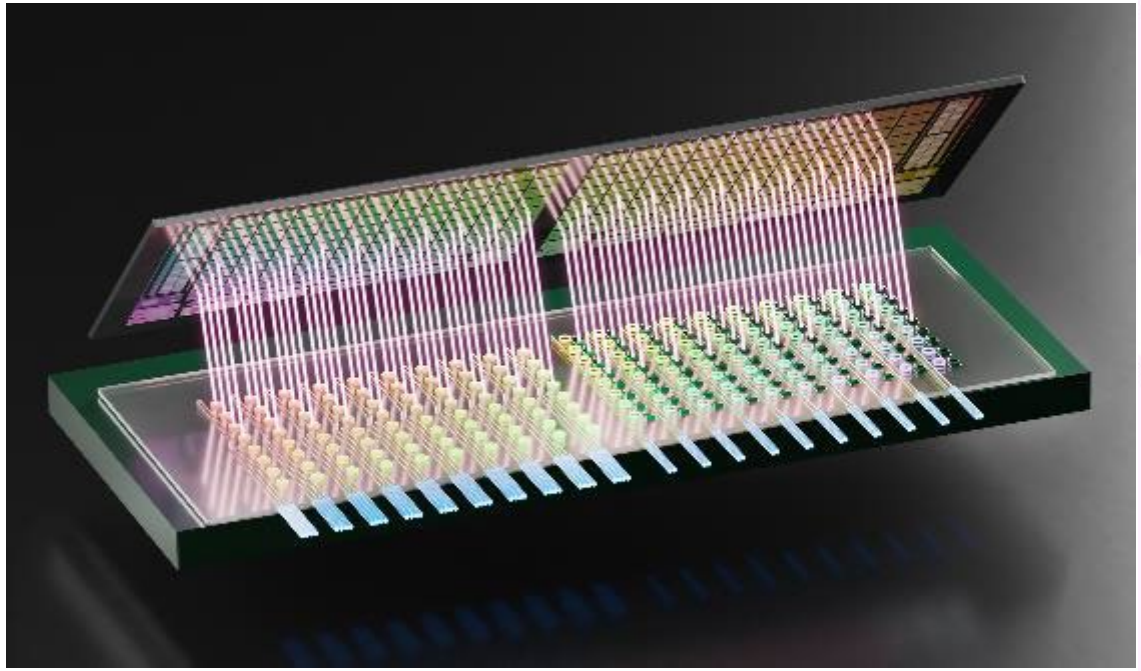


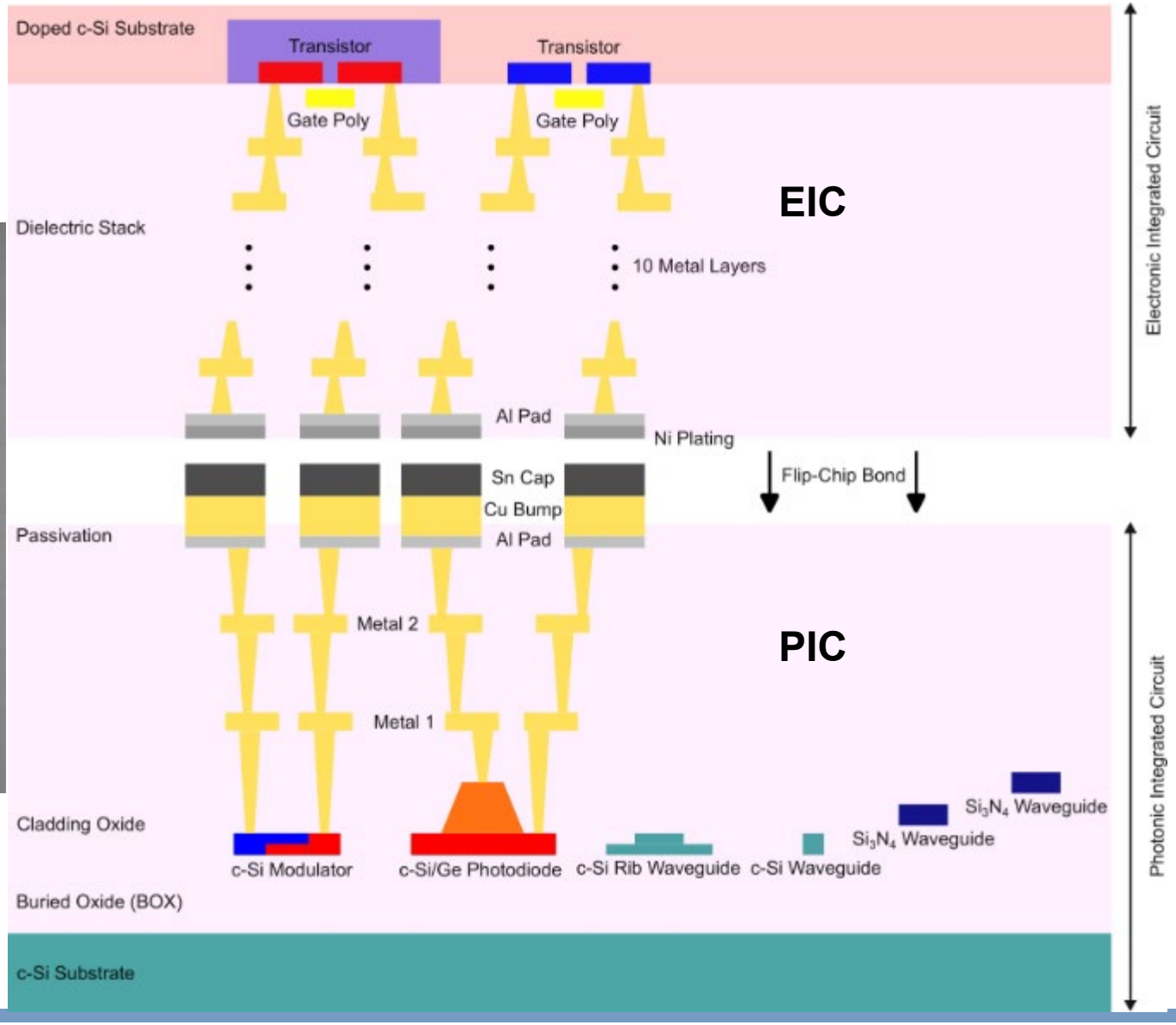
Microdisk Filter

- Fab-robustness w/o inner sidewall
- Modulation efficiency custom vertical junction
- Fabricated FSR closely match design values



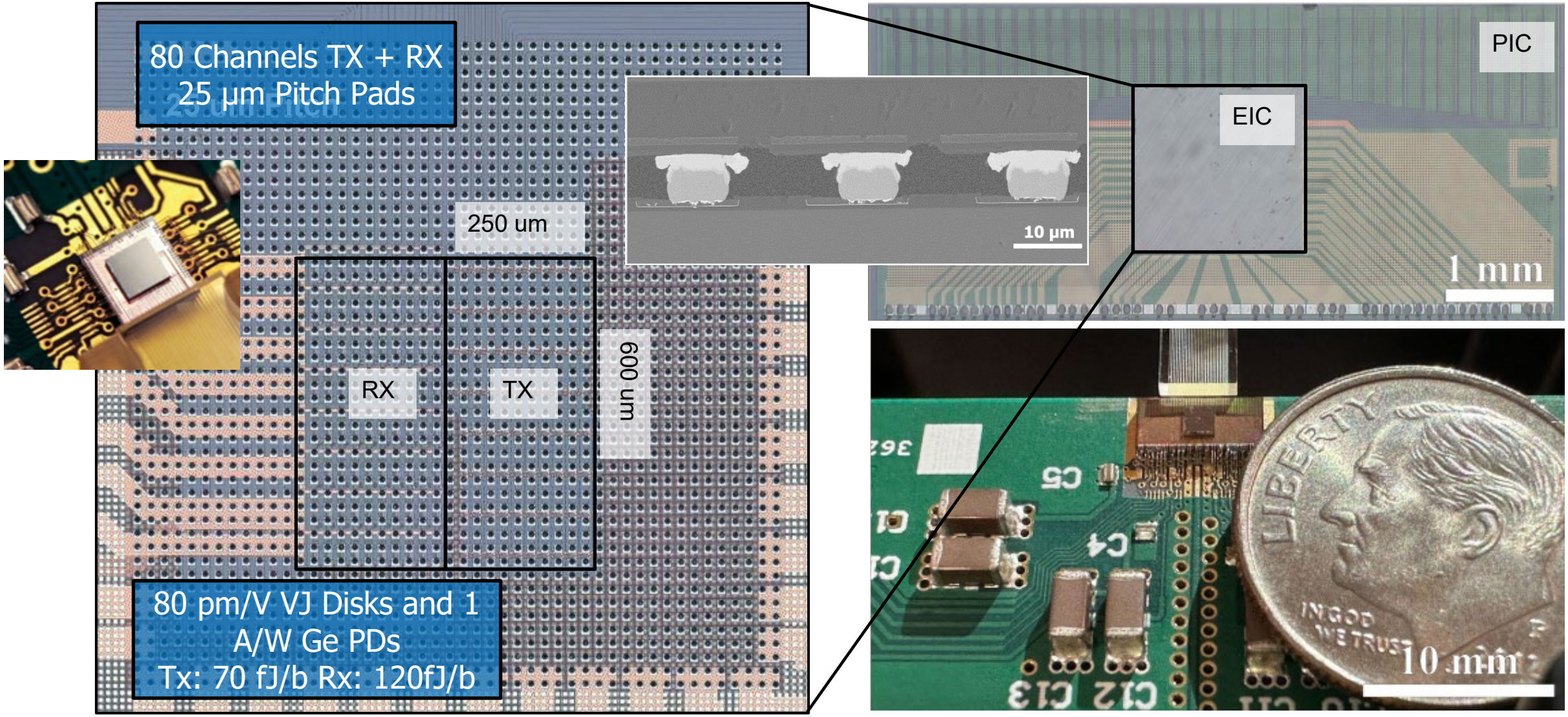
3D Photonic IO





S. Daudlin, K. Bergman et al., “3D photonics for ultra-low energy, high bandwidth-density chip data links,” Nature Photonics (March, 2025)

Fully Assembled High Density 5.3 Tb/s/mm^2 MCM



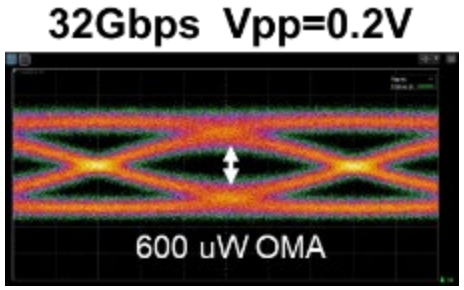
Energy Breakdown

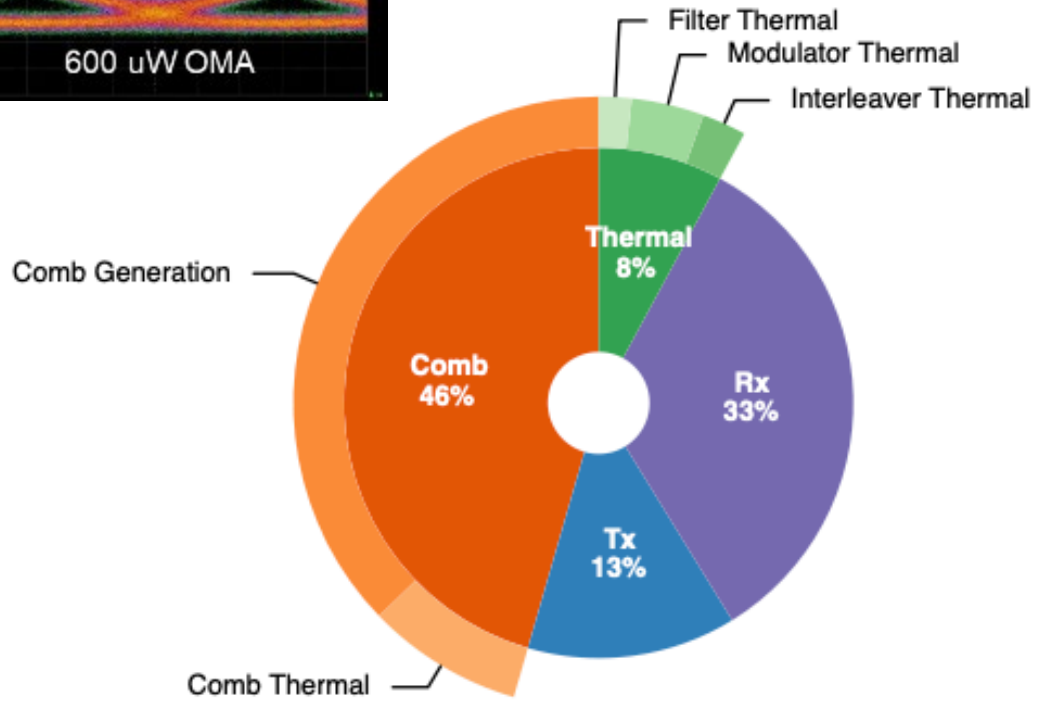
- **2.048 Tb/s** per fiber: 64 λ 32 Gbps/channel
- **> 5.4 Tb/s/mm edge** bandwidth density

Energy Breakdown

	Component	Energy [fJ/b] w/o undercut	Energy [fJ/b] w/ undercut
Comb*	Comb Generation	112.3	**112.3**
	Comb Thermal	24.9	**24.9**
EIC	Tx Driver	40.0	**40.0**
	Rx TIA	100.4	**100.4**
PIC	Interleaver Thermal	35.2	**7.0**
	Modulator Thermal	58.0	**11.6**
	Filter Thermal	26.0	**5.2**
	Total	396.7	**301.4**

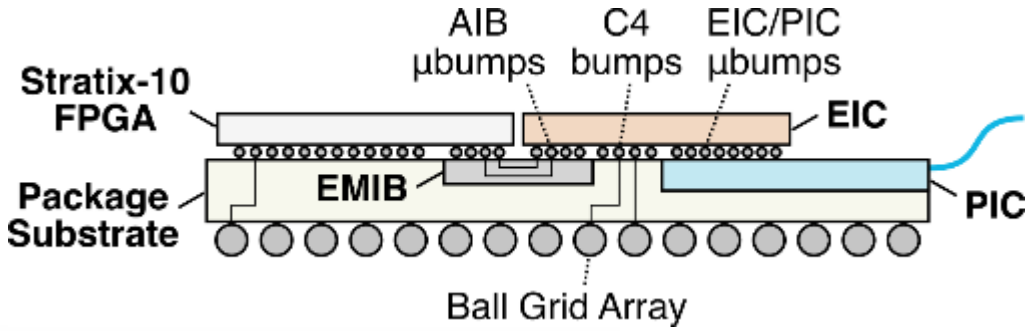
* Assuming 15 % overall comb WPE.

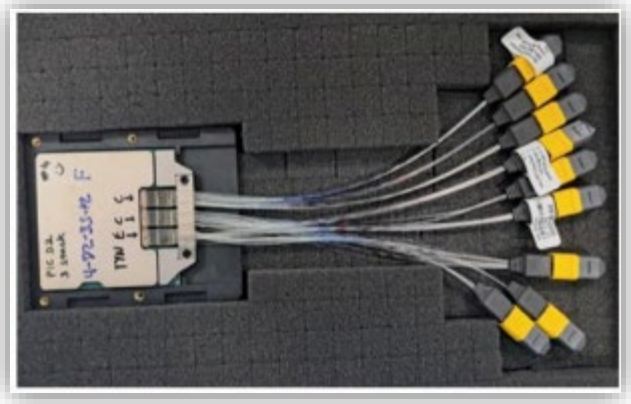




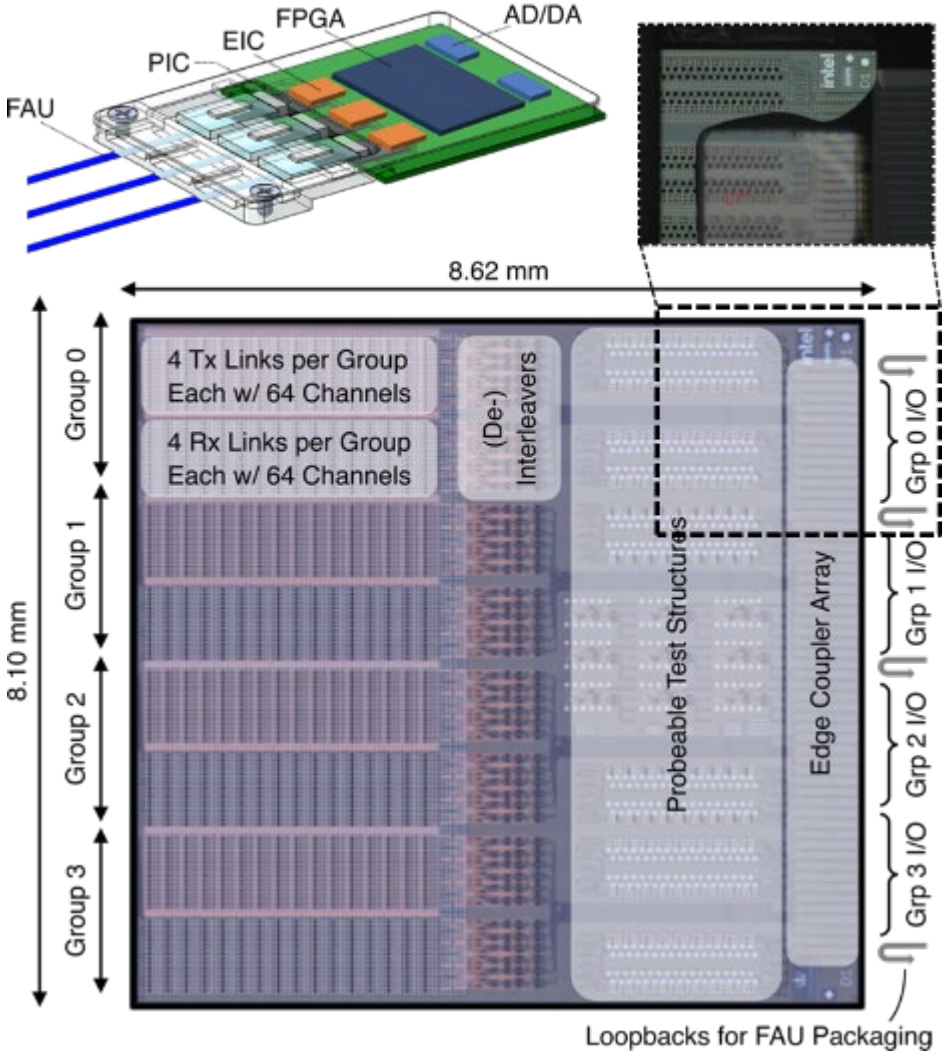
Bidirectional 96 Tb/s Multi-Chip Package

- 1024 disk modulators and 1024 filters integrated per PIC
- 3 PICs per package, 3D integrated EICs
- 16 Gbps/channel ⇒ >16 Tbps across ~8 mm shoreline
- 2 Tbps/mm (4 Tbps/mm bidirectional) bandwidth density





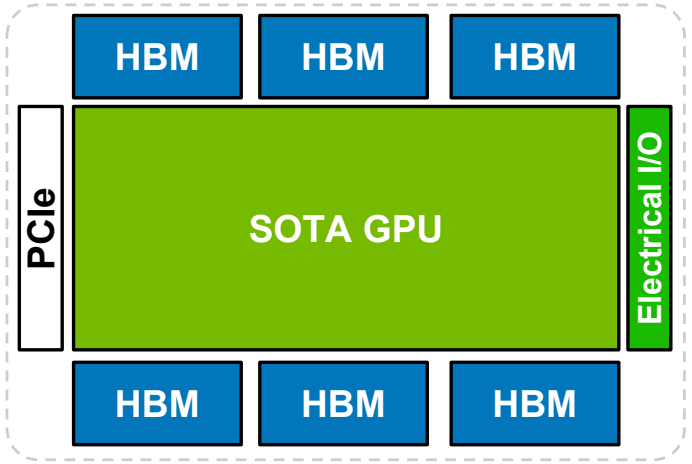
Multi-chip package (MCP) consisting an FPGA and **96 Tb/s** bidirectional bandwidth achieved by 6 optical I/O chiplets.

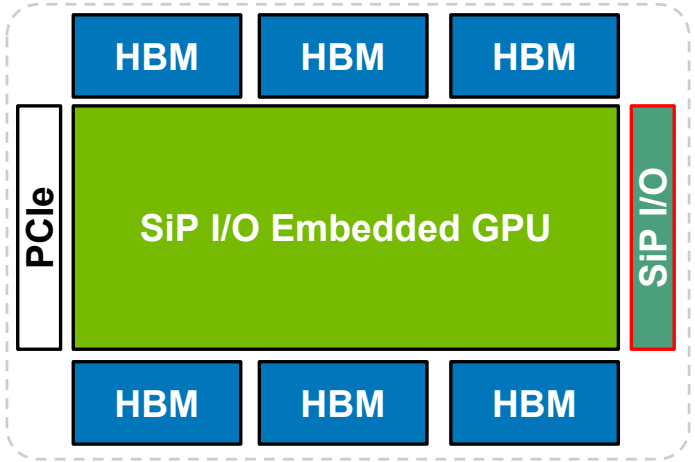


Yuyang Wang *et al. IEEE CICC*, 2024, Wang, Y. et al. IEEE T-CPMT (2024)

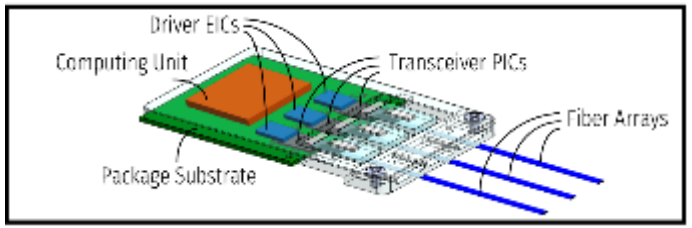
Expanding the Memory Pooling Design Space

Compute die's *shoreline width* is used a critical resource.

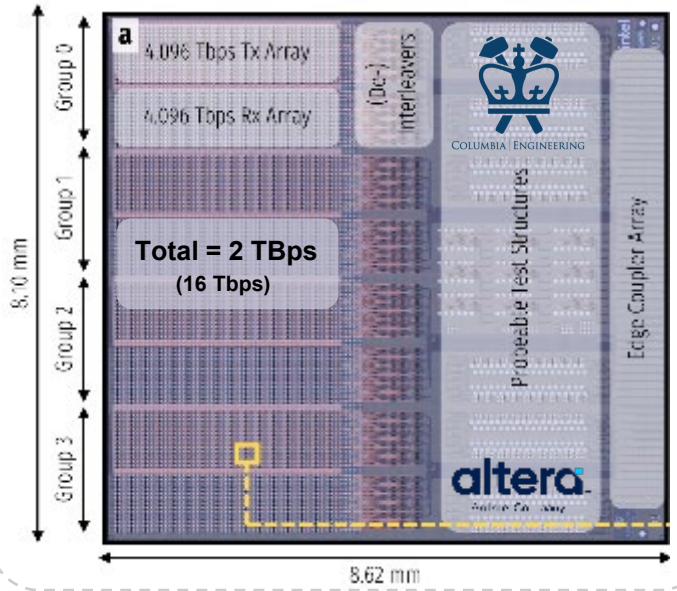


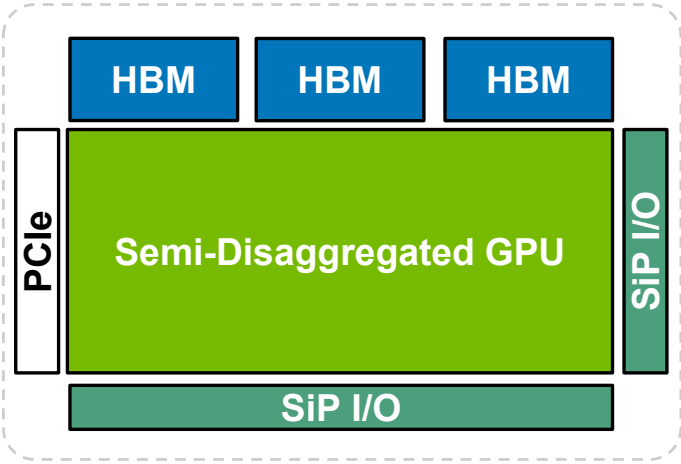


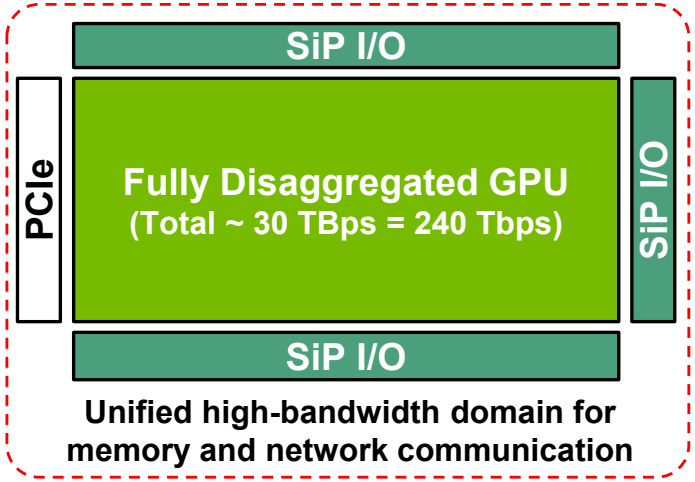
Conceptual Embedded SiP Chip I/O



Designed and fabricated transceiver PIC

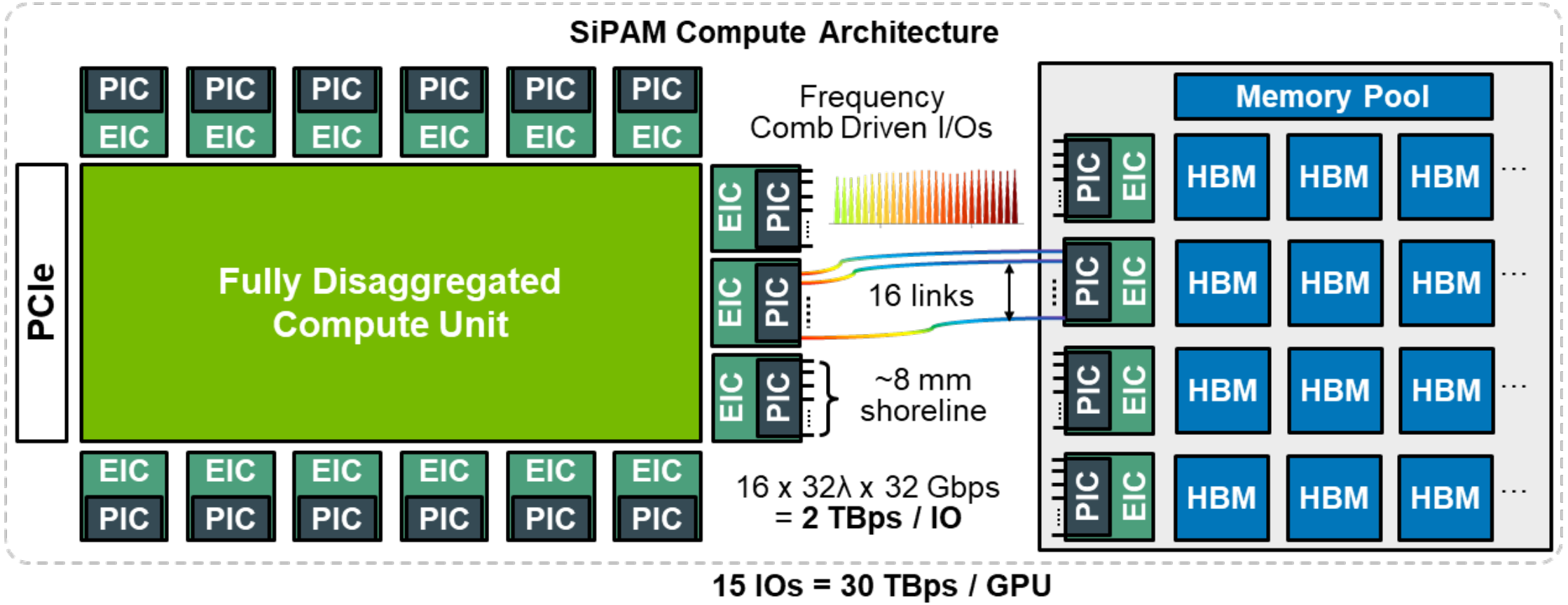




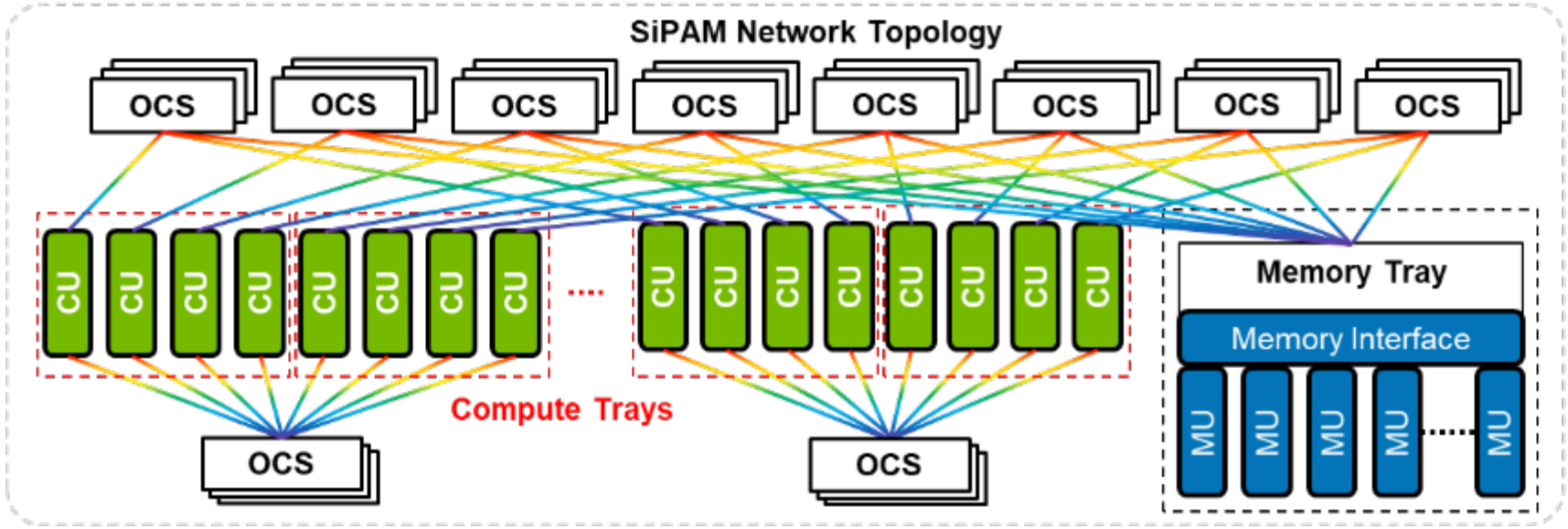


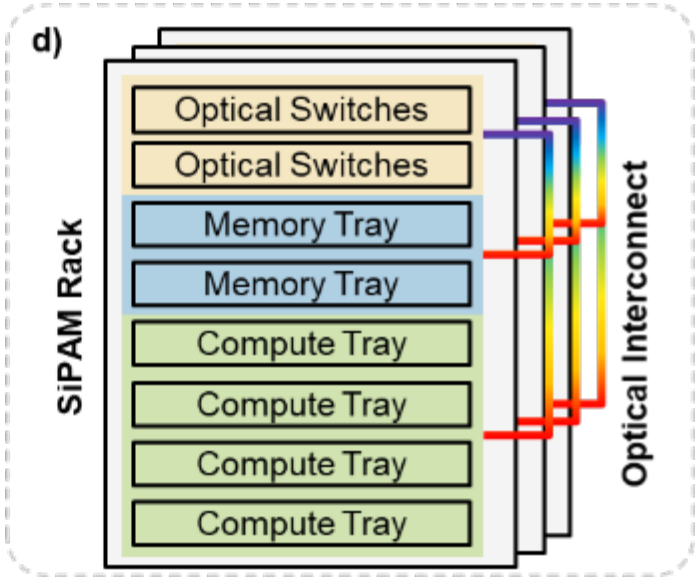
Unified high-bandwidth domain for memory and network communication

SiPAM: Silicon Photonic Accelerated Memory-Pooling



SiPAM: Silicon Photonic Accelerated Memory-Pooling





- Each SiP I/O can be **flexibly allocated** for high-speed memory access or network communication.
  - One-shot reconfiguration per workload

Evaluation Setup

Hardware – Nvidia GPU Based

Single CU	FP16 TFLOPs	Mem Cap (GB)	Mem BW (TBps)
Nvidia A100	312	40	1.5
Nvidia H100	1000	80	3
Nvidia B100	3500	192	8
SiPAM*	3500	Up to 720	Up to 30

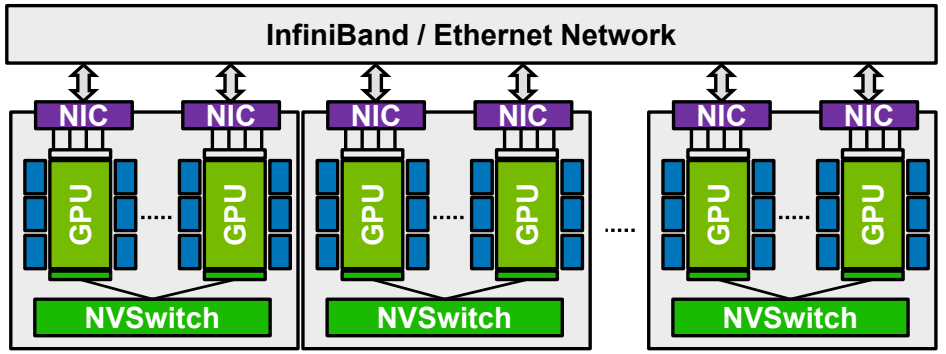
❖ **Cluster Size:** up to 1024 GPUs

Cluster of 1024 CUs	FP16 PFLOPs	Mem Cap (TB)	Mem BW (PBps)
Nvidia A100	320	41	1.5
Nvidia H100	1024	82	3.1
Nvidia B100	3584	197	8.2
SiPAM*	3584	Up to 737	Up to 31

* Assuming B100 as CU and HBM3E as MU

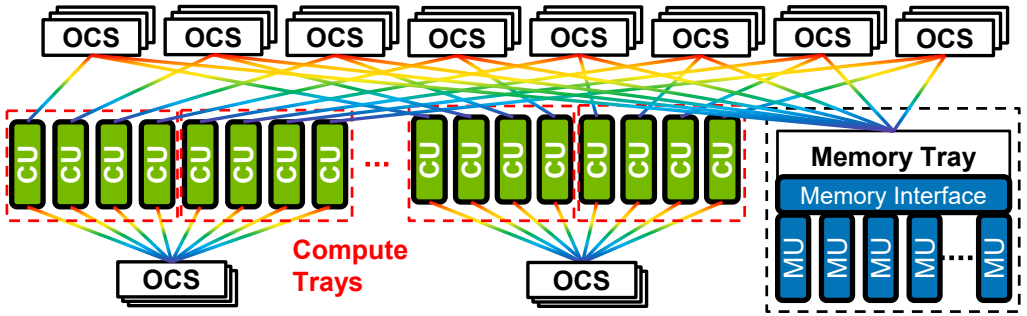
Baseline Configuration:

• NIC: up to **800 Gbps** / GPU



❖ NVL Domain: up to 72 GPUs

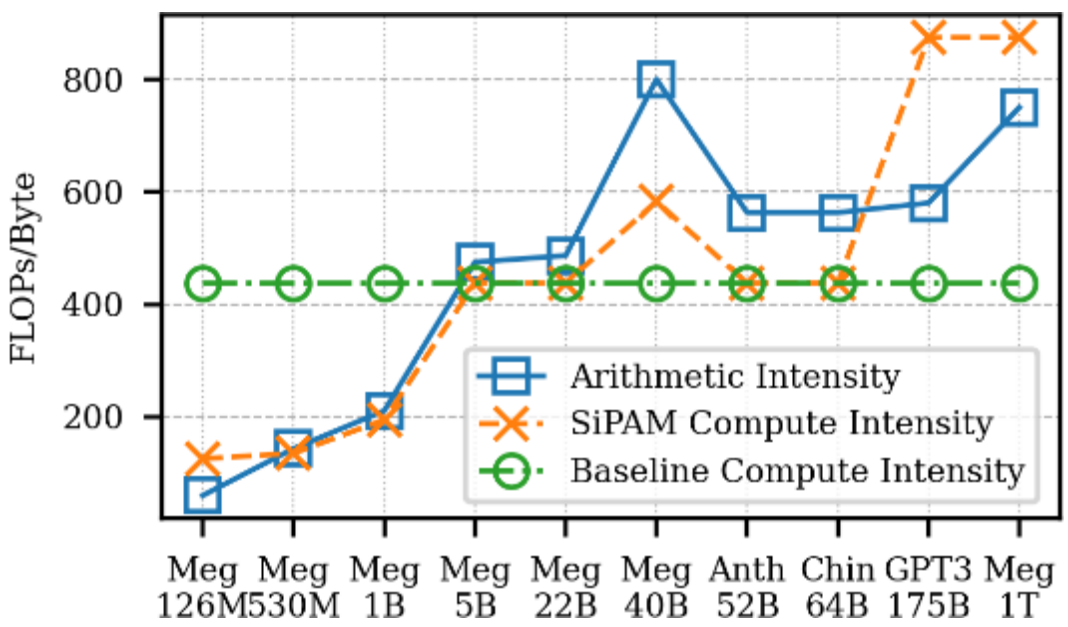
SiPAM Configuration:

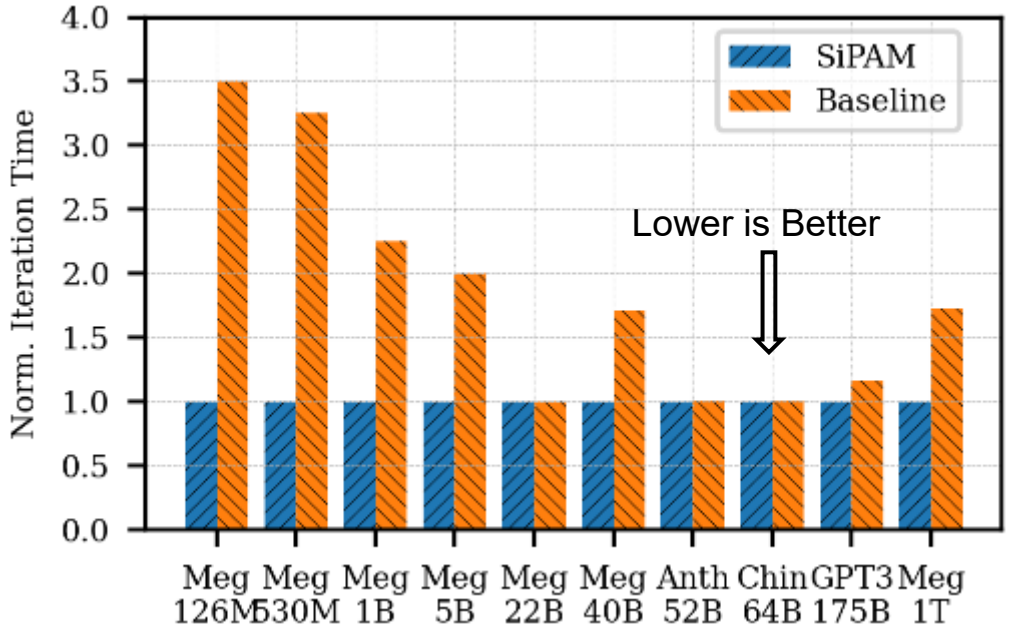


[Calculon] Isaev, Mikhail, et al. "Calculon: a methodology and tool for high-level co-design of systems and large language models." SC 2023.

Simulation Results - Training

- **Workloads:** Megatron-126M/5B/22B/40B/1T, Anthropic 52B, Chichilla-64B, GPT3-175B (**Training**)
- **Baseline:** Up to 256 B100 GPUs each with fixed 192 GB HBM memory @ 8 TBps total memory bandwidth
- **SiPAM:** Up to 256 GPUs, with compute, memory bandwidth, and capacity optimized based on each workload



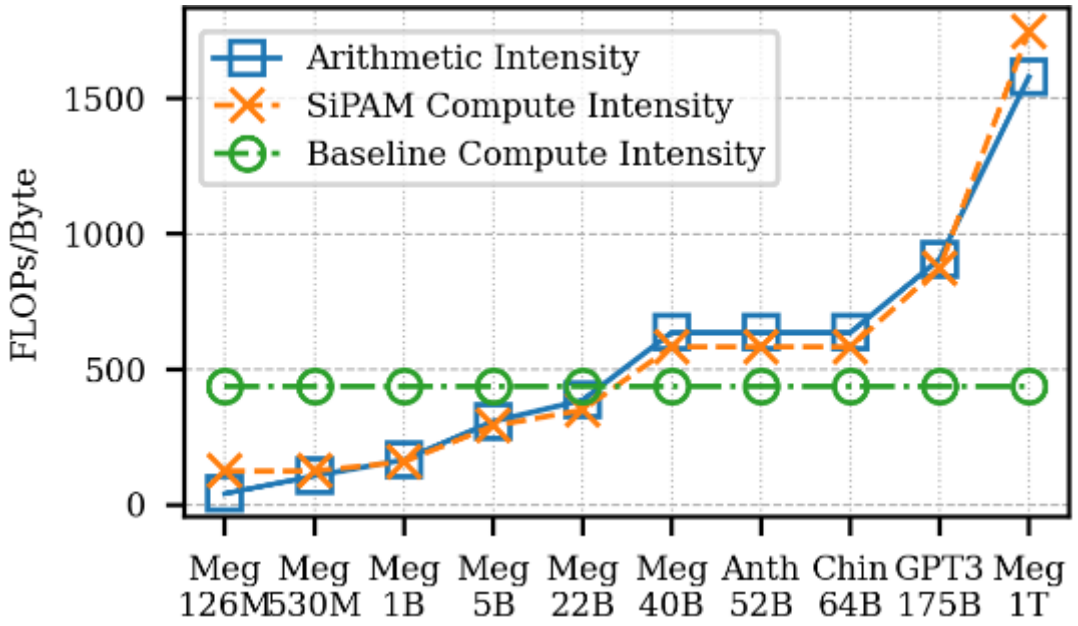


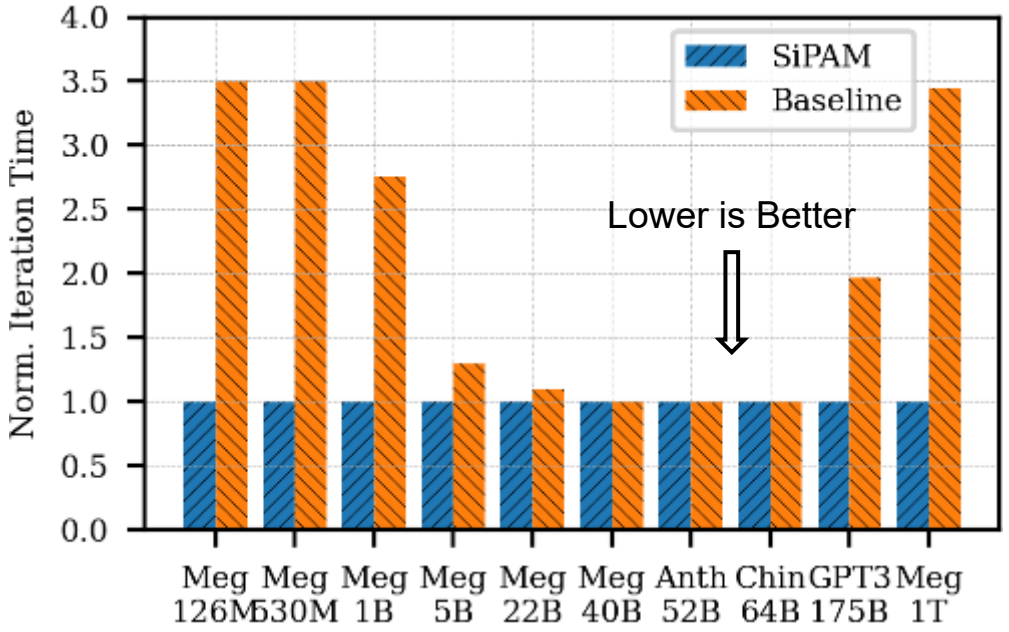
a) SiPAM tracks arithmetic intensity closely, while the baseline remains constant

b) SiPAM improves training time by up to **3.5x**

Simulation Results - Inference

- **Workloads:** Megatron-126M/5B/22B/40B/1T, Anthropic 52B, Chichilla-64B, GPT3-175B (**Inference**)
- **Baseline:** Up to 64 B100 GPUs each with fixed 192 GB HBM memory @ 8 TBps total memory bandwidth
- **SiPAM:** Up to 64 GPUs, with compute, memory bandwidth, and capacity optimized based on each workload

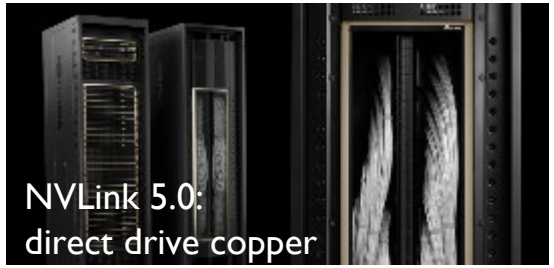




a) SiPAM tracks arithmetic intensity closely, while the baseline remains constant
b) SiPAM improves inference time by up to **3.5x**

Pushing the Limits of AI Systems with Embedded Photonics

Nvidia's GB200 NVL72

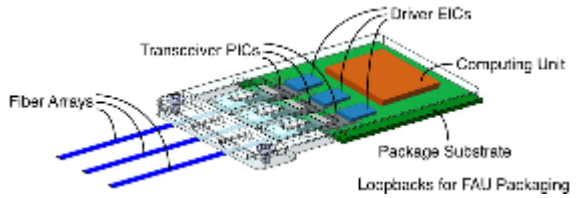


- Copper-based electrical links
- Limited scale-up domain (72)
- Nvidia's higher per-chip compute is constrained by its electrical interconnects when scaling for aggregate compute performance.

Huawei CloudMatrix 384

- Linear Pluggable Optical cables
- Increased scale-up domain (384)
- Higher total compute power due to larger scale-up domain (with lower per-GPU compute)
- Lower power-per-bit

Embedded photonics is the solution!



- Much higher bandwidth density (> 4 Tbps/mm)
- Much lower energy consumption (sub-pJ/bit)
- Distance agnostic scaling (>1000 GPUs in scale-up)

Scaling AI Networking Infrastructure

Compute / Memory / Interconnect Comparison*

Chip-Level	Nvidia GB200	Ascend 910C	Embedded Photonics
TFLOPs	2,500	780	-
HBM Capacity (GB)	192	128	Up to 720 GB
HBM Bandwidth (TBps)	8	3.2	> **240 Tbps**
Scale Up Bandwidth (Tbps)	7.2	2.8	
Scale Out Bandwidth (Tbps)	0.4	0.4	

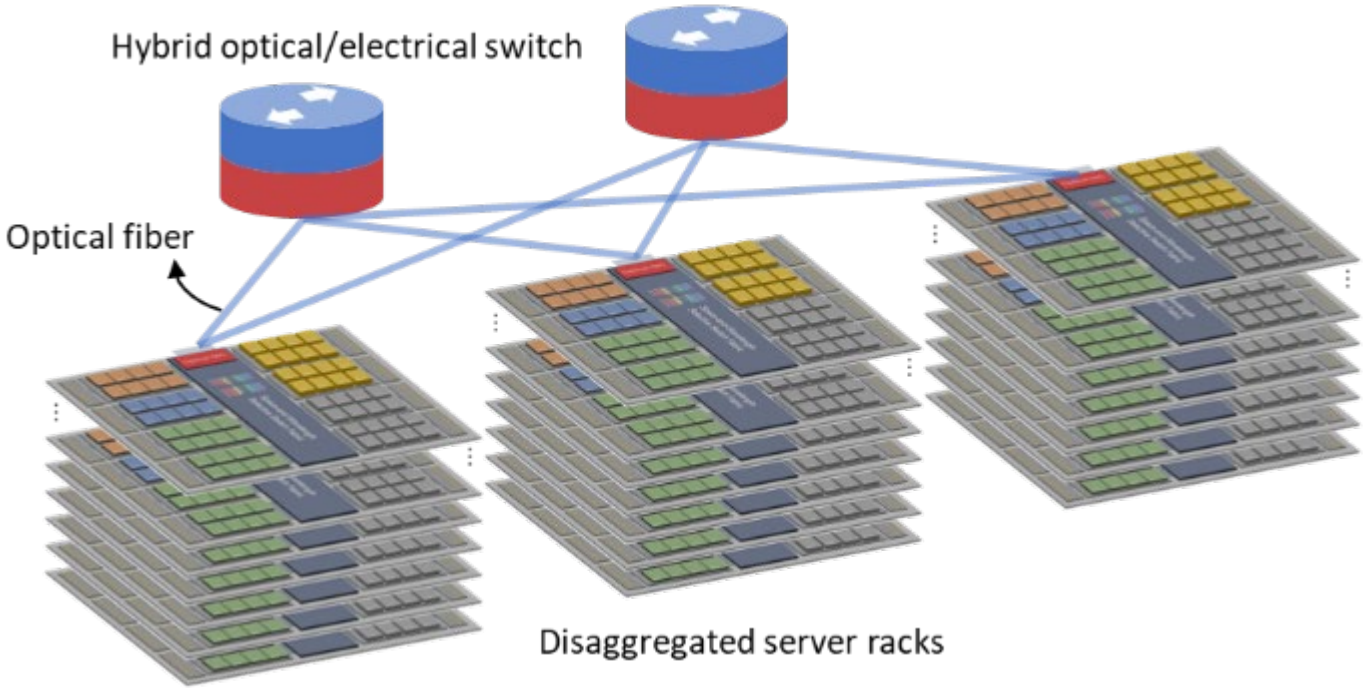
System-Level	GB200 NVL72	CLoudMatrix 384	Embedded Photonics
# Compute Units	72	384	> 1000
PFLOPs	180	300	-
All-In System Power (kW)	145	599	-
HBM Capacity (TB)	13.8	49.2	Up to **276.5 TB**
HBM Bandwidth (TBps)	576	1,229	Up to **11,520 TBps**
Scale Up Bandwidth (TBps)	64.8	134.4	
Scale Out Bandwidth (TBps)	3.6	19.2	

* Numbers based on SemiAnalysis report: https://semianalysis.com/2025/04/16/huawei-ai-cloudmatrix-384-chinas-answer-to-nvidia-gb200-nvl72/
** Estimated based on 384 compute-unit scale-up domain size

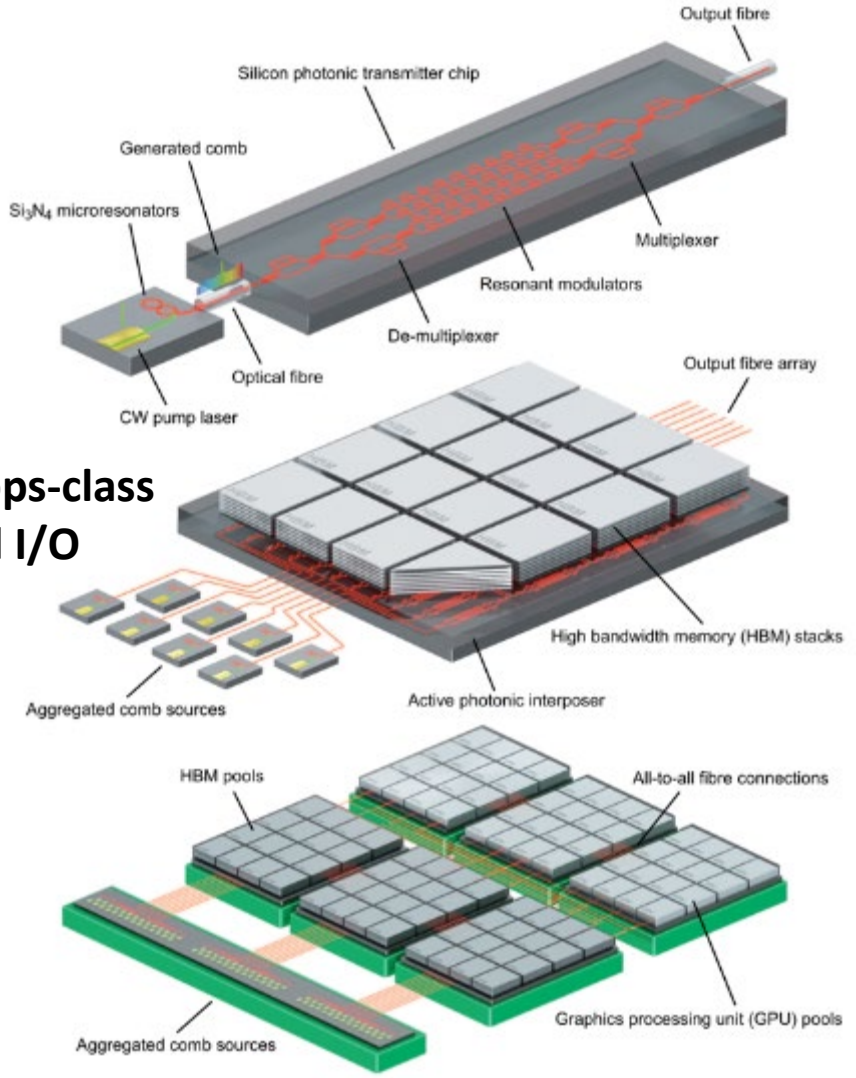
27

Scalable Energy Efficient AI Photonic Architectures

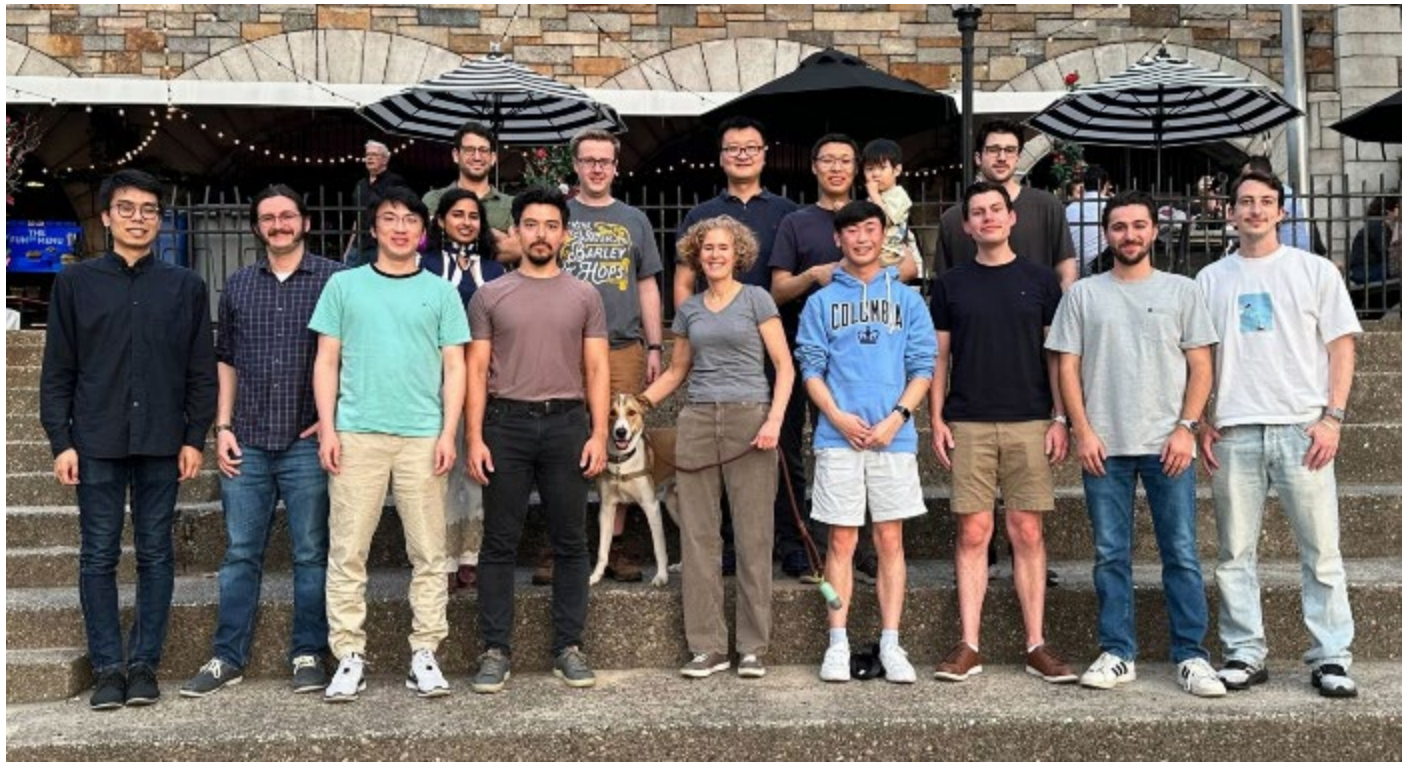
- System scalability with photonic connectivity
- Flexible, composable to workloads

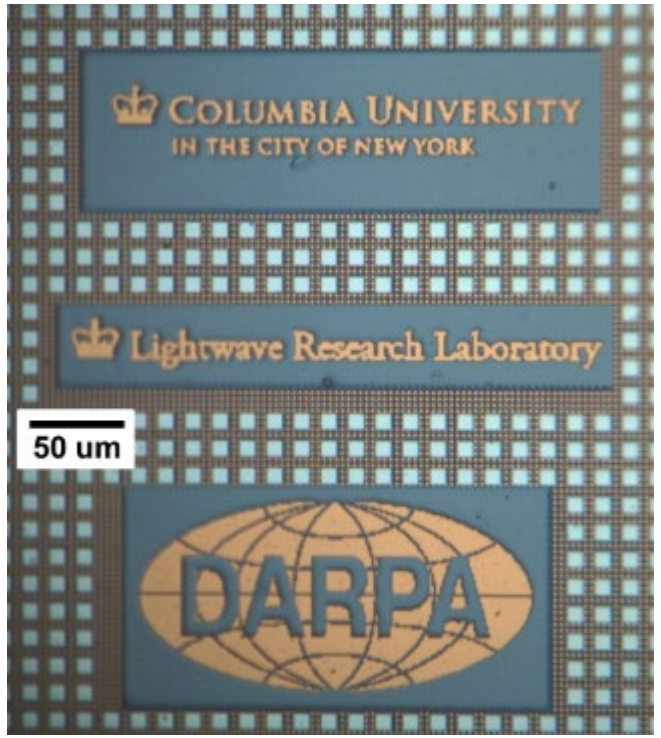


100 Tbps-class optical I/O



Source: A. Rizzo *et al.*, Nature Photonics, 2023




COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK
Lightwave Research Laboratory
50 um
DARPA

CUbiC


arm
IBM
ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™
SK hynix
Micron
BOEING
EMD ELECTRONICS
HRL LABORATORIES
tsmc
qorvo
NSF
intel
DARPA
GlobalFoundries™
Raytheon Technologies
SAMSUNG
MEDIATEK