# Petascale Photonic Connectivity for Energy Efficient AI Computing

**Keren Bergman**
Department of Electrical Engineering
Columbia University, New York, NY

2023 MulticoreWorldX

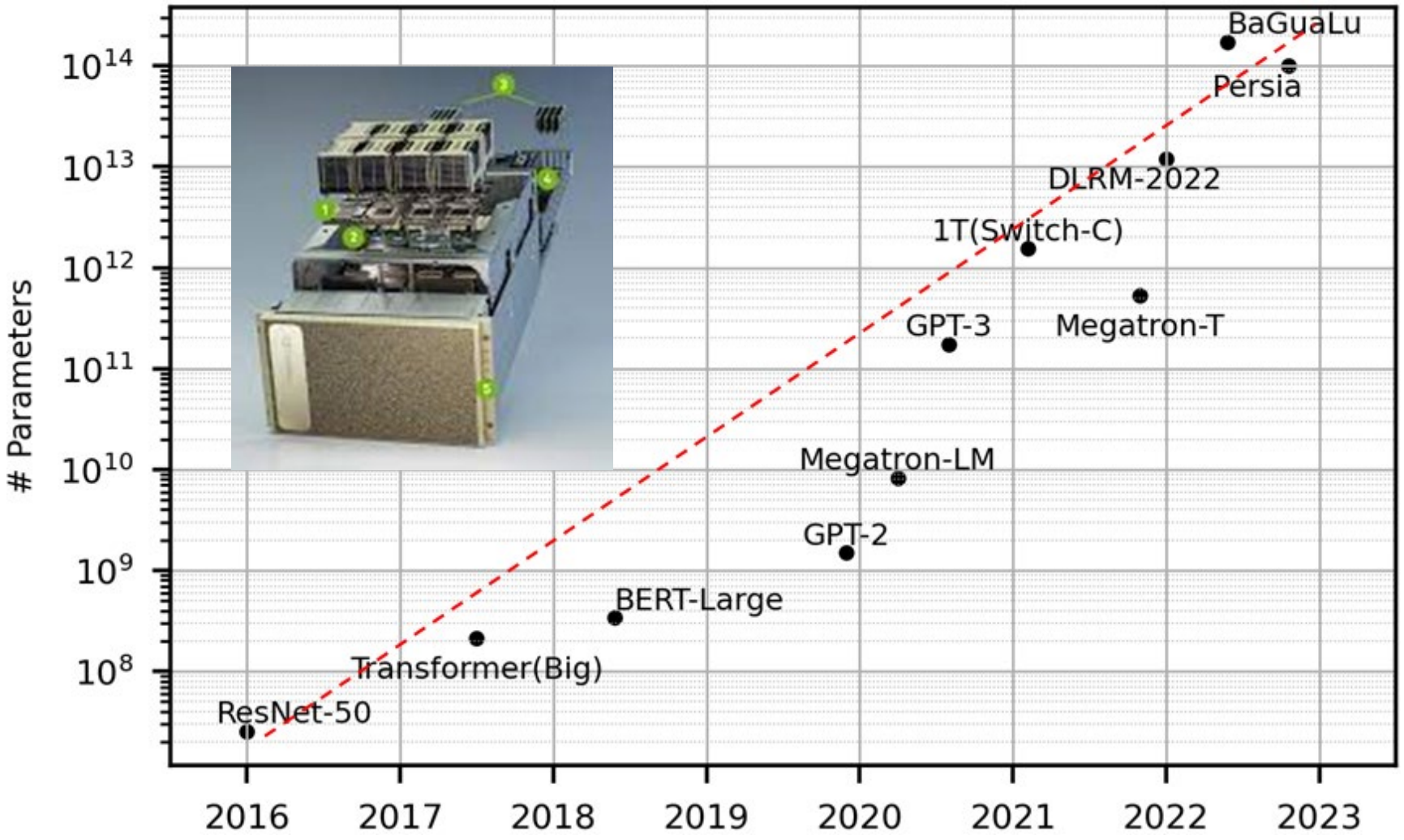
COLUMBIA ENGINEERING
The Fu Foundation School of Engineering and Applied Science

# Intro – Keren Bergman

- BS EE Bucknell Univ; MS and PhD EE MIT
- Professor and faculty director CNI, Columbia University
- DOE ASCR; DARPA Exascale Computing Initiative
- ISC 2022 Technical Chair
- Silicon photonics 300mm foundry, Leadership Council
- Lead PI: ARPA-E ENLITENED Program; DARPA ERI - PIPES
- Director, SRC JUMP 2.0 Center for Ubiquitous Connectivity (CUbiC) – DARPA, 15 industry partners

- Fellow IEEE, Optica

# AI Applications Driving Ever Larger Models for Deep Learning

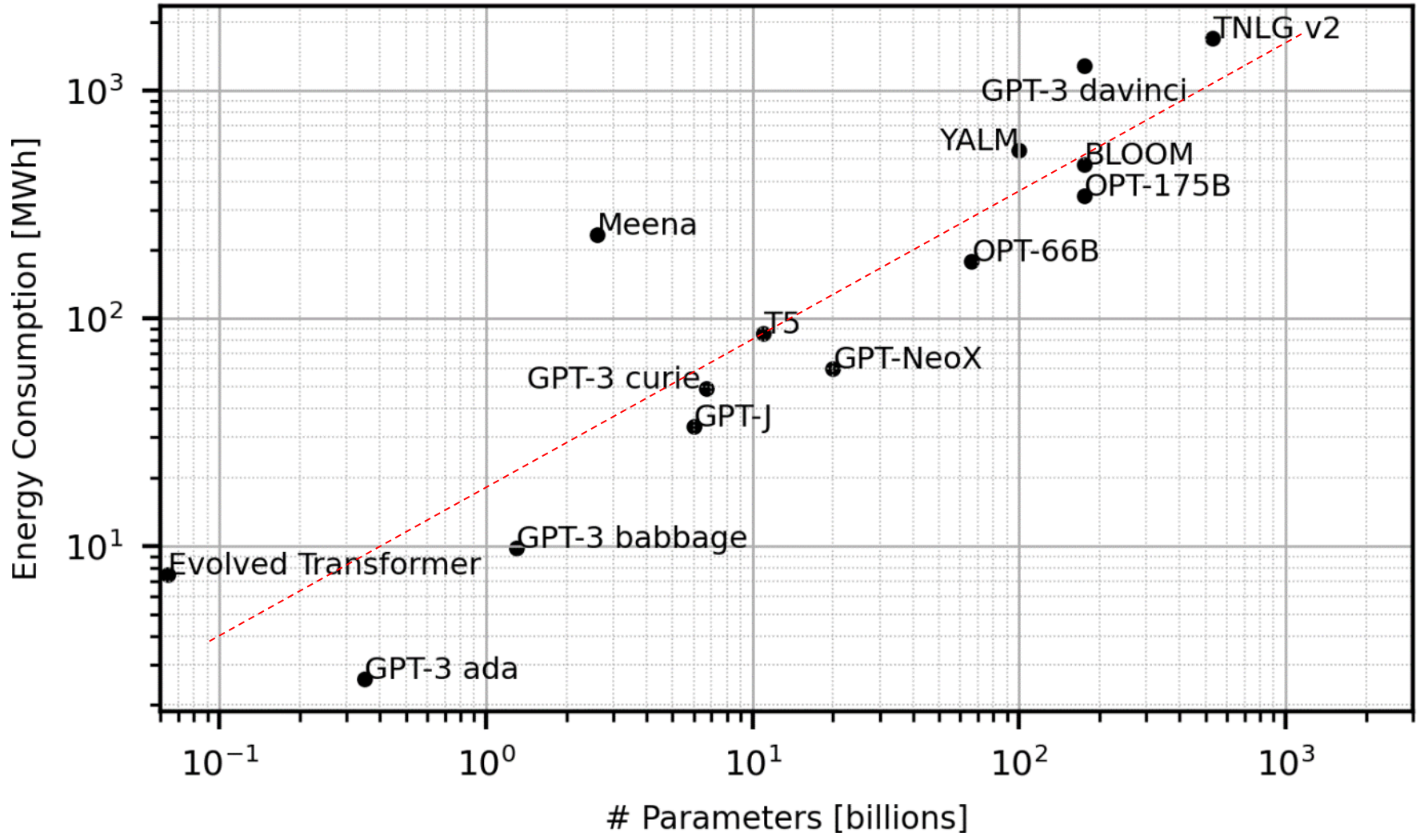



Model sizes increased
> 6 orders of magnitude
in 6 years

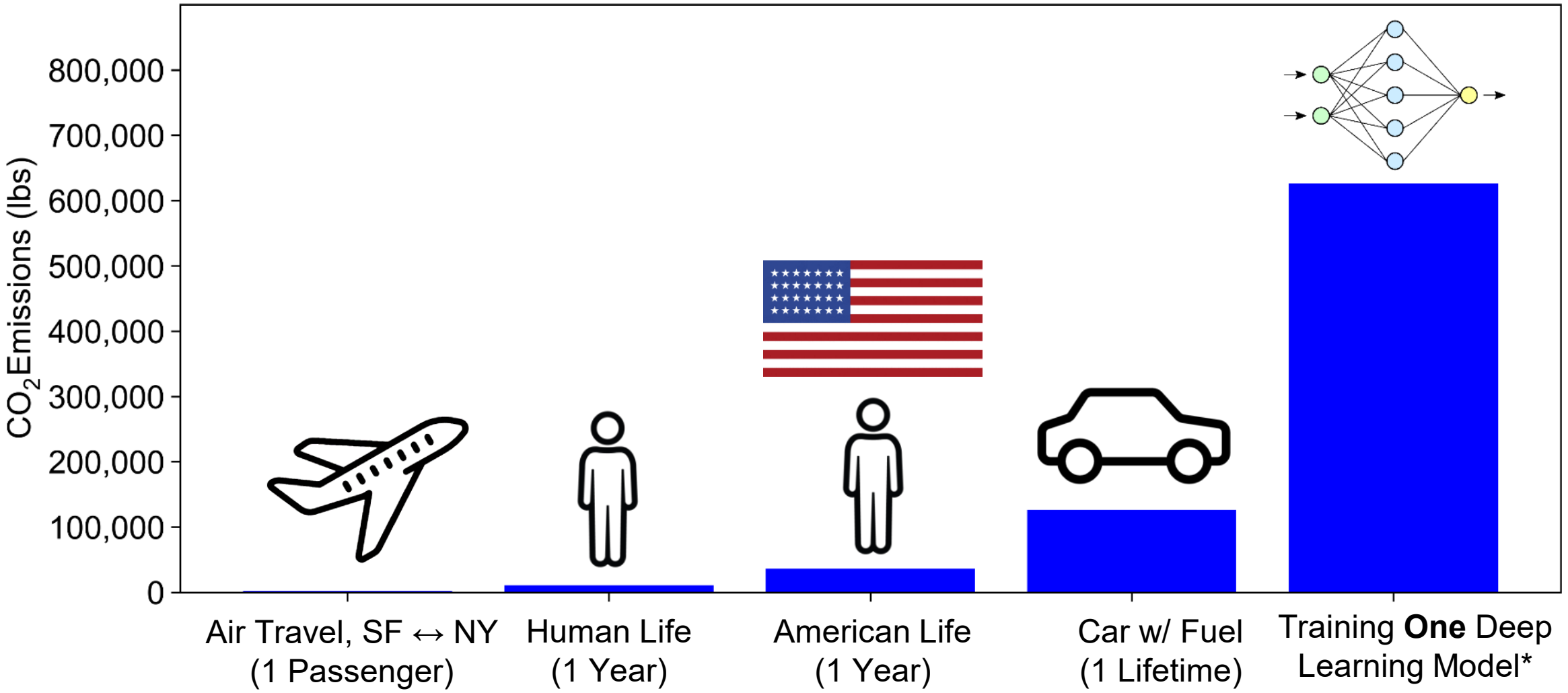
> 10 Trillion parameters

Exceeds memory capacity of any single computing unit

# Per-Training Energy Consumption



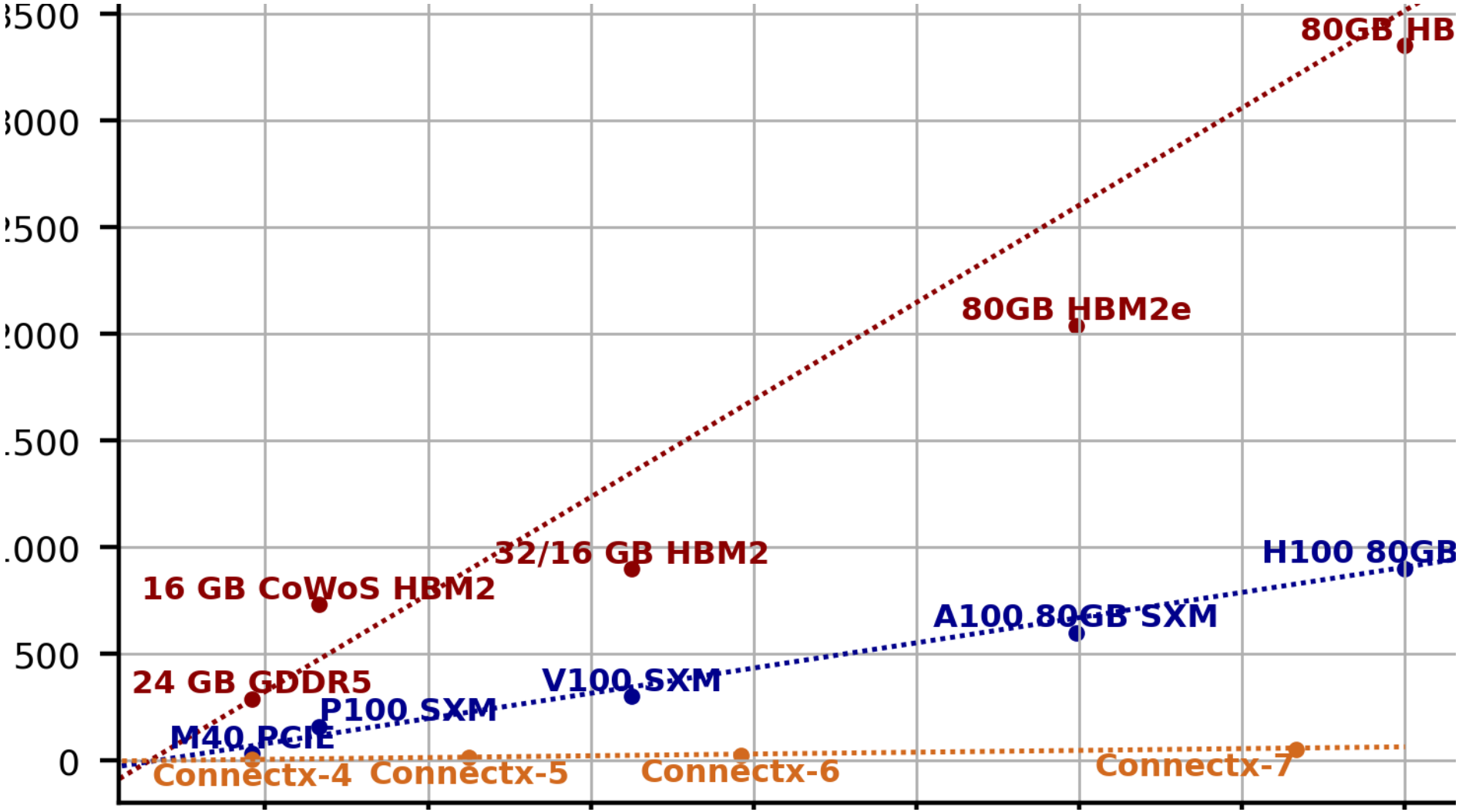
# ML Training – Workloads Energy Consumption



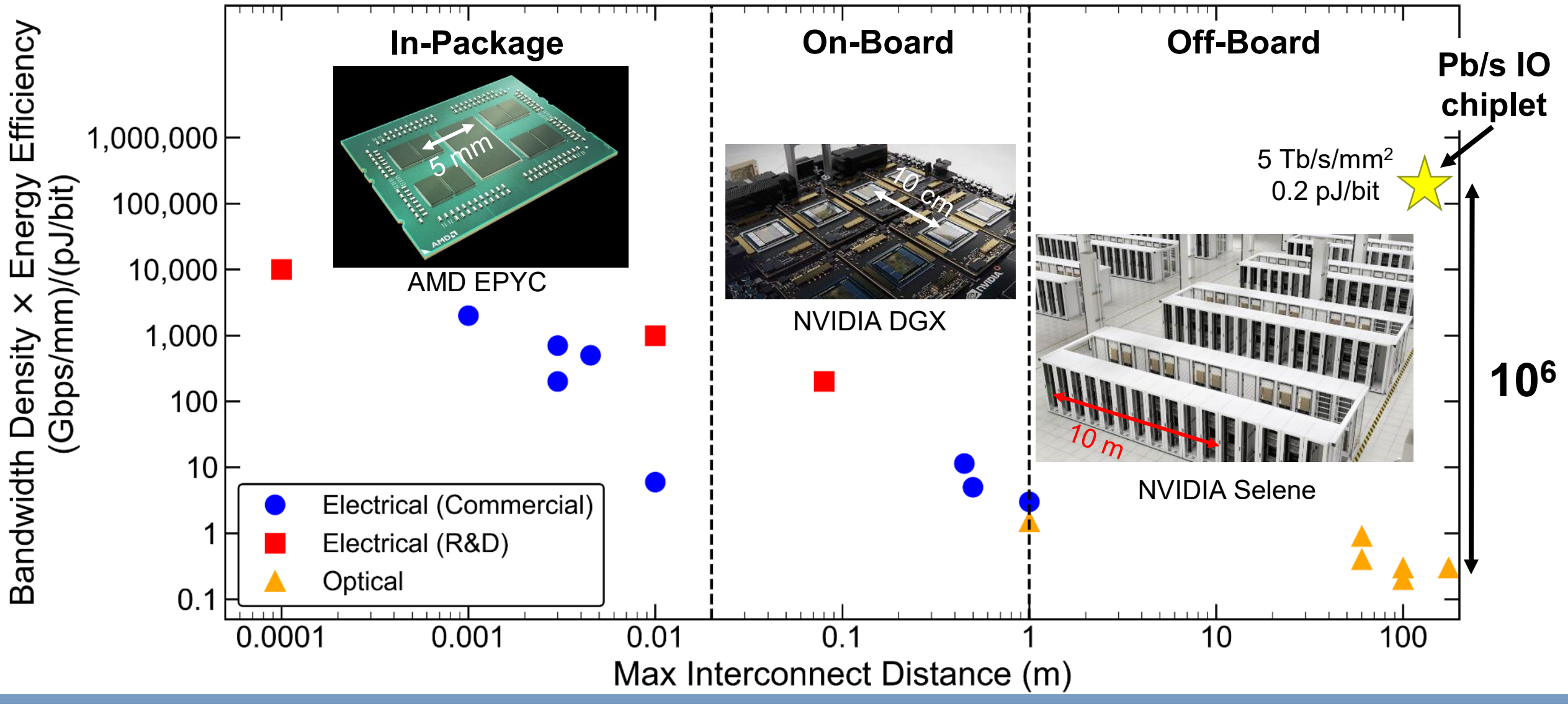

*State-of-the-art neural architecture search, trained on 8 NVIDIA P100 GPUs (1,515 W), ~ 656,000 kWh [see *arXiv:1906.02243* for full assumptions]

Adapted from E. Strubell, A. Ganesh, A. McCallum, *arXiv:1906.02243* (2019)

# Distributed Deep Learning: Communications Bottleneck



# Bringing Photonics to the Chip

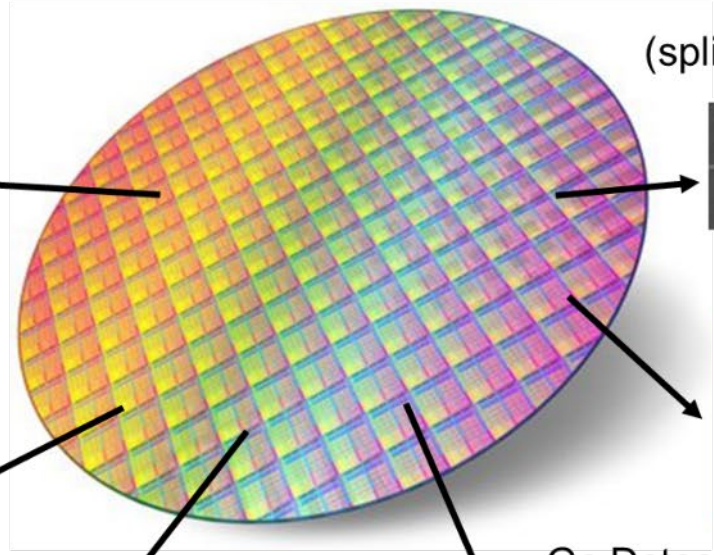



Adapted from Gordon Keeler, DARPA

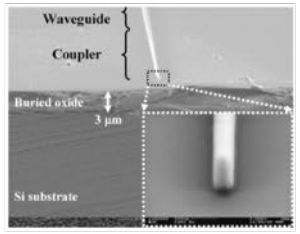
# Silicon Photonics Fabrication

GlobalFoundries™ intel AIM PHOTONICS

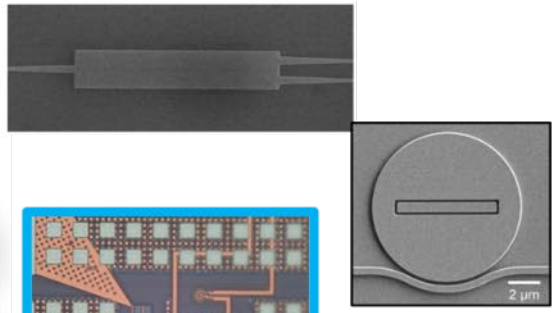
300 mm SOI Wafers

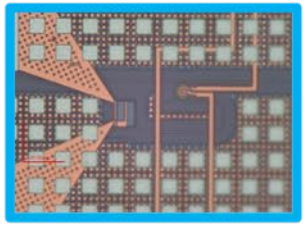


Low Loss
Chip Coupling

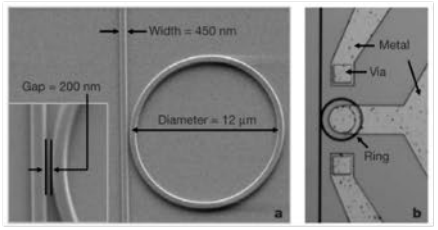



High Performance Passives
(splitters, filters, polarization control)

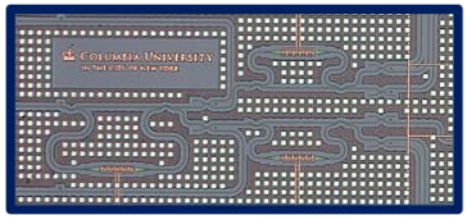


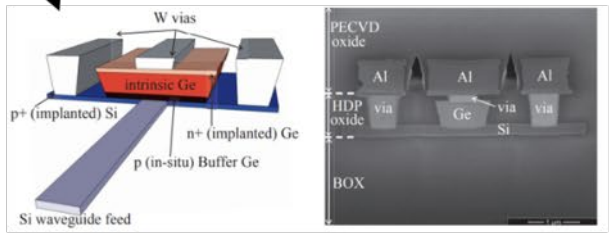
High Speed Modulators




Wavelength
interleaving

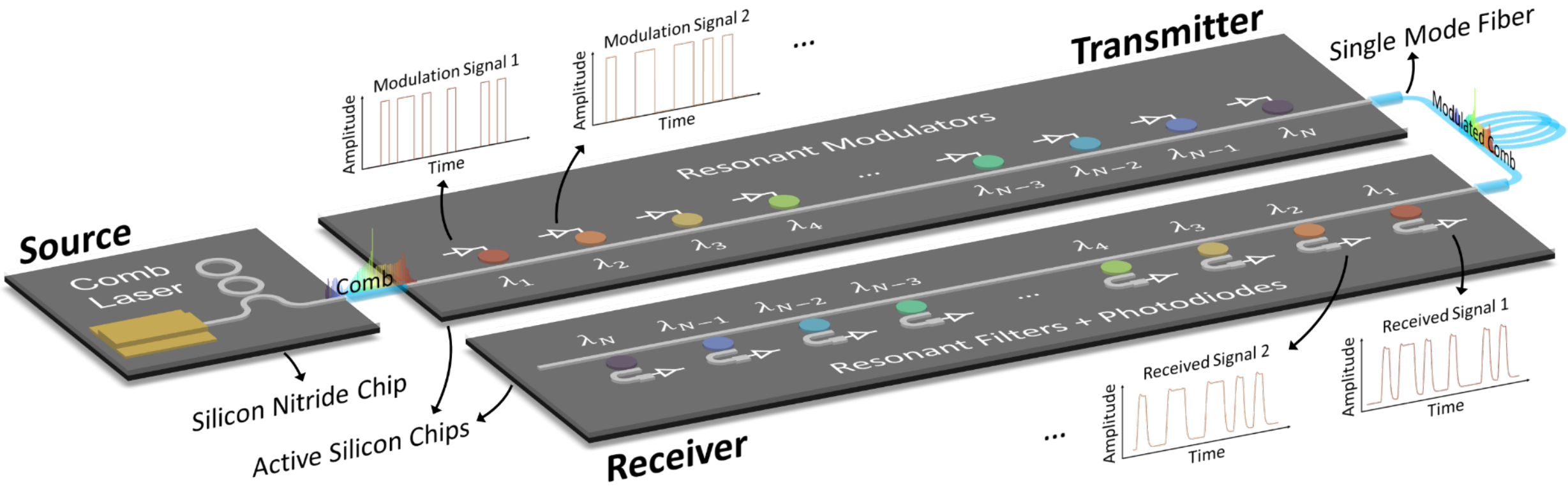


Ge Detectors




# Photonics = Massive Parallelism in the Wavelength Domain

**Frequency Combs: Multi-Tb/s per Single Link**



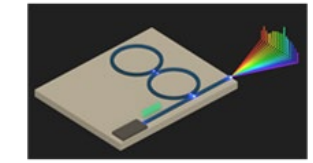
# Approach to reaching multi-Tbps IO and sub-pJ/b

**Key Technical Innovations:**

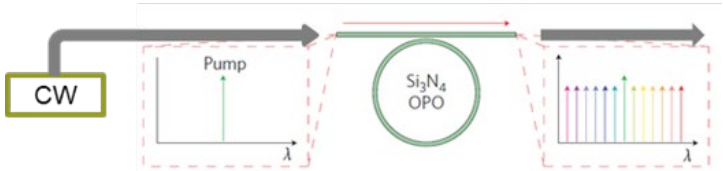
- Embrace extreme parallelism:
  - Ultra-dense channels generated by >250 wavelengths (WDM) comb source
  - Each wavelength channel modulated at modest data rates for minimizing energy consumption
  - SERDES-*less* operation; energy/bandwidth density co-optimization

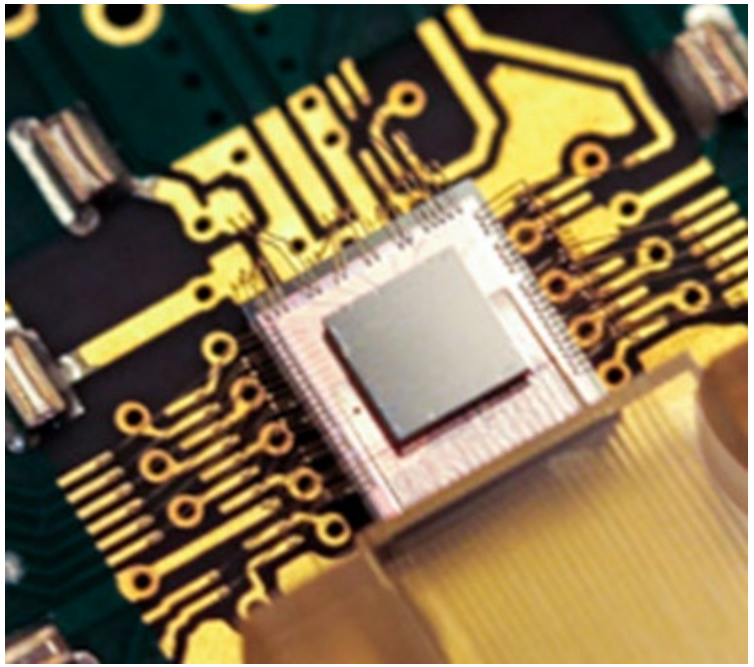
- Scalable link architecture:
  - Co-design with broadband comb source
  - Multi-FSR operation regime

- Reduction of thermal energy consumption:
  - Photonics *robust* to fabrication variations
  - Engineered for athermal operation
  - Wafer scale undercut for increased efficiency

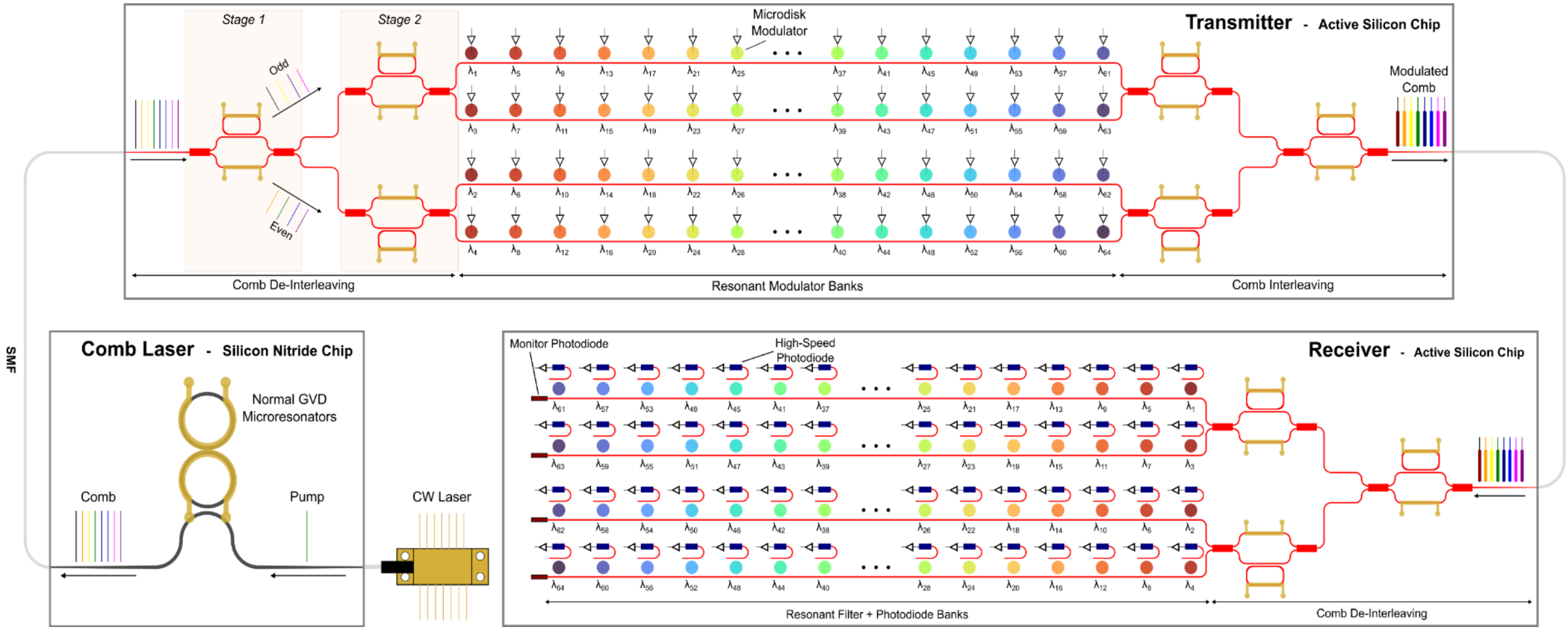


Frequency comb

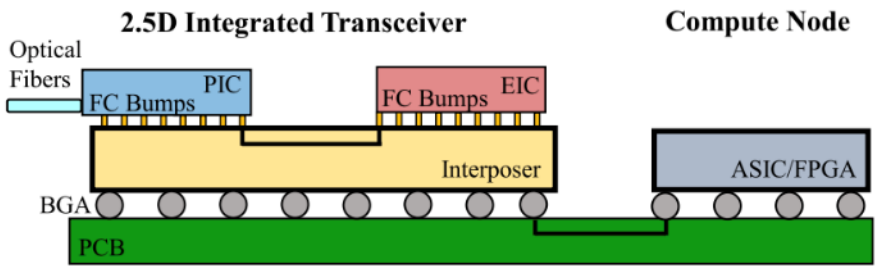


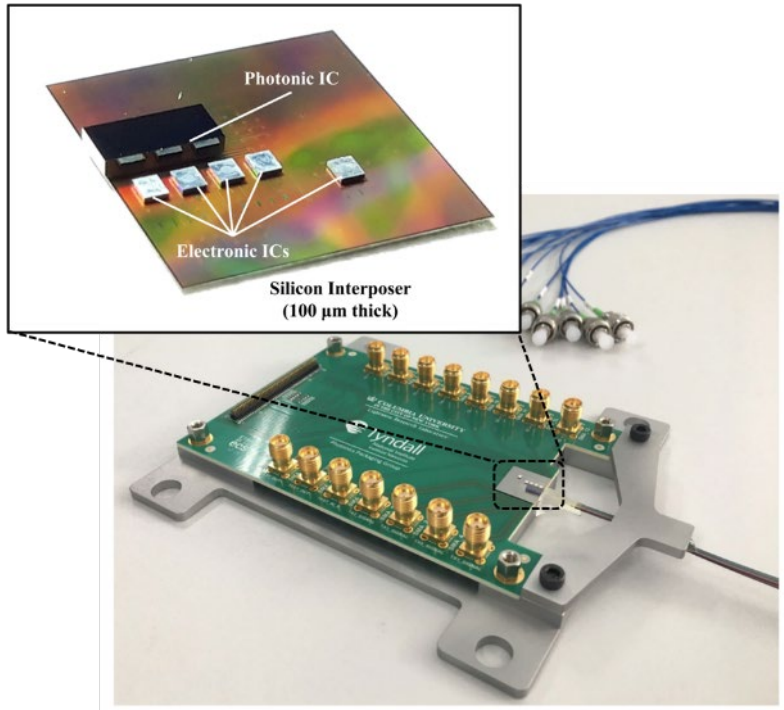



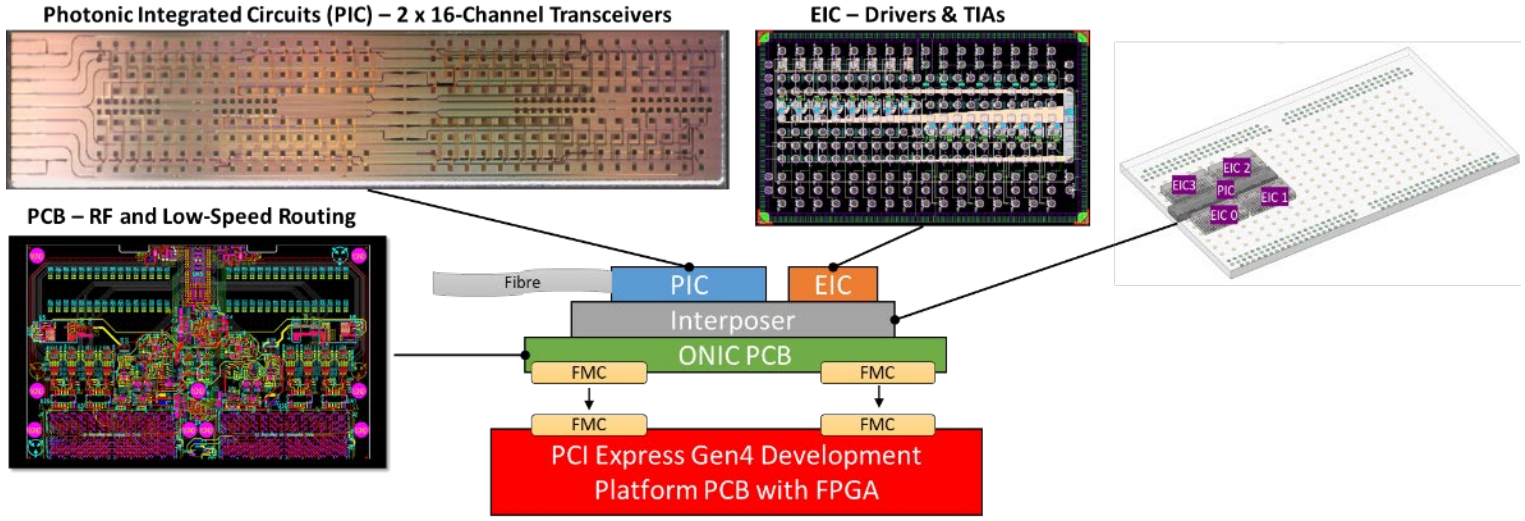
# Scalable Photonic Link Architecture

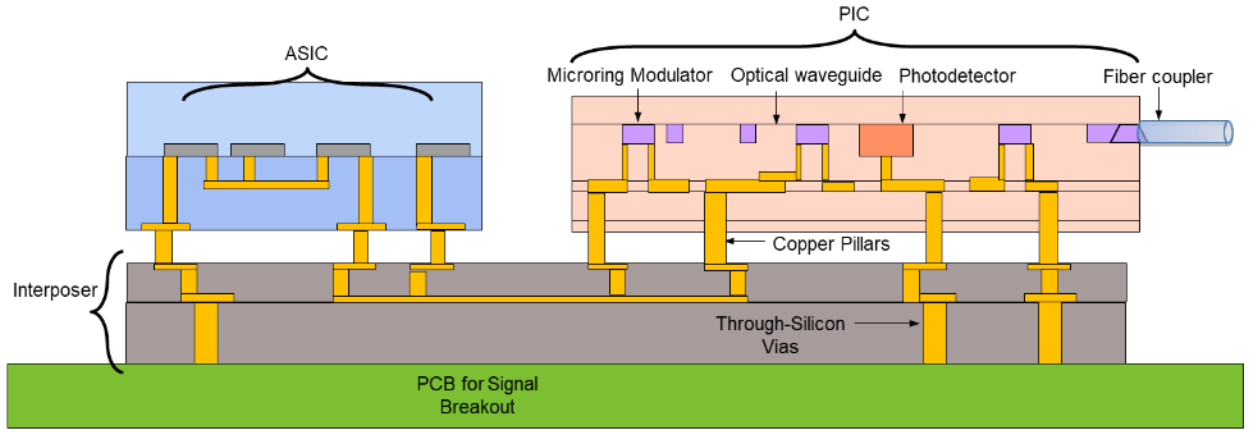


# 2.5D High-Density Packaging – Enables Systems Exploration

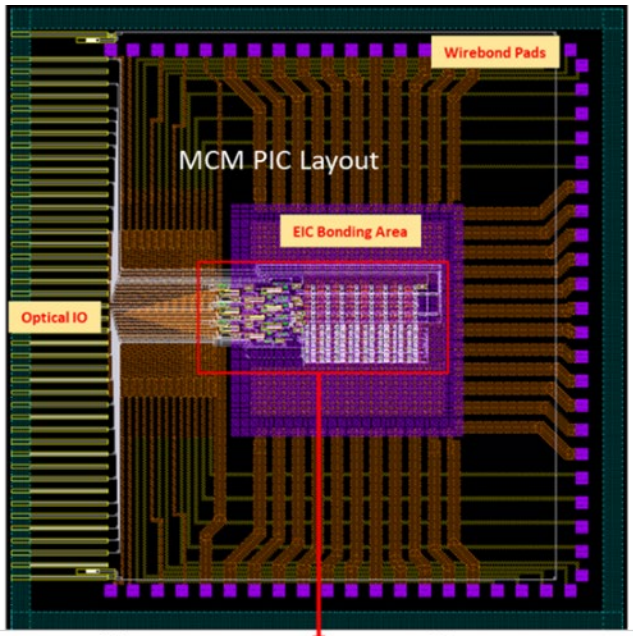


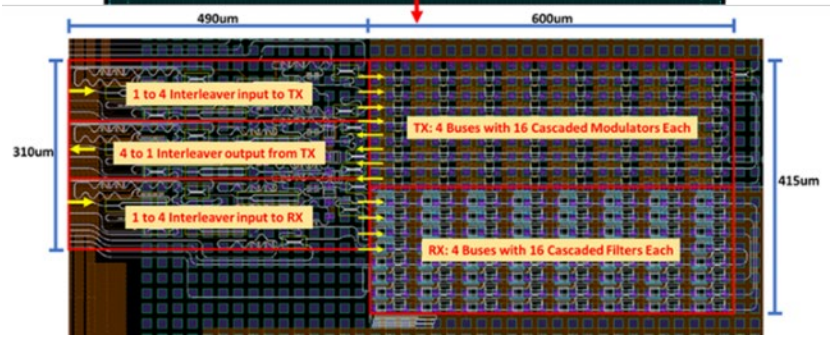



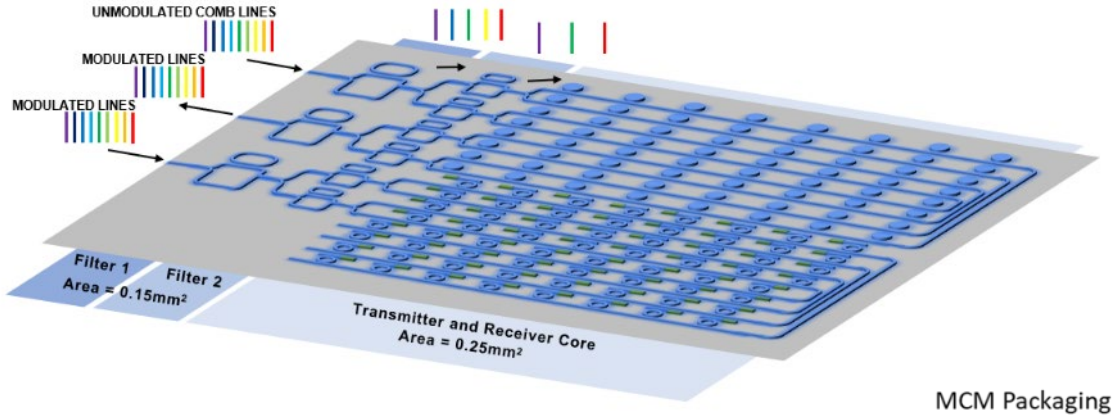



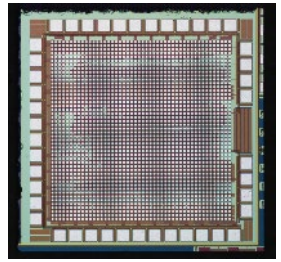



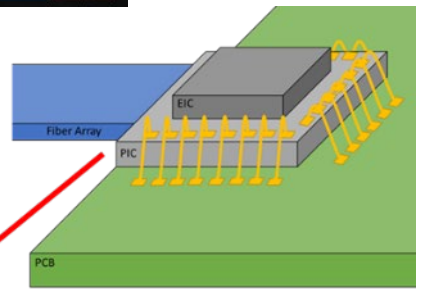
# 3D Integration to Realize Bandwidth Density



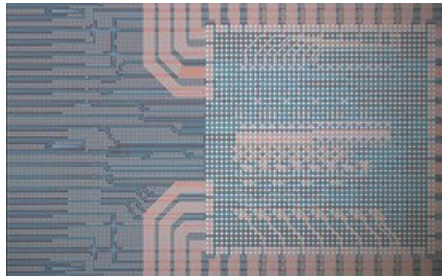



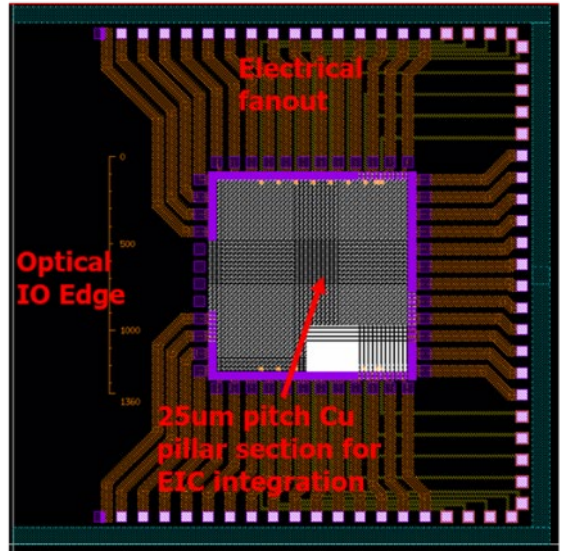



- Transceiver: 600um x 415um = 0.25mm$^2$
- Interleavers: 490um x 310um = 0.15mm$^2$
- Bandwidth density:

  2Tbps / 0.4mm$^2$ = 5Tbps/mm$^2$



MCM Packaging Diagram

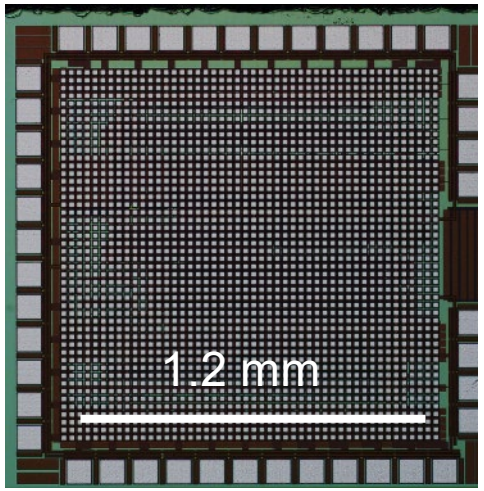




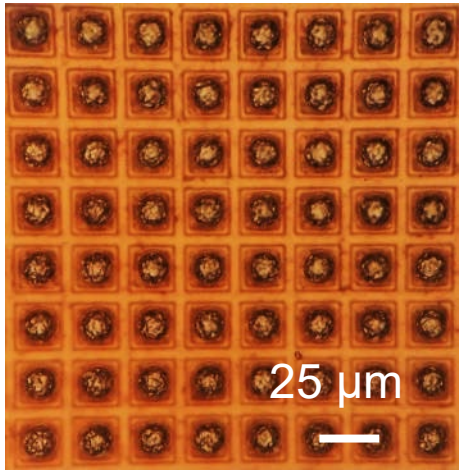
13

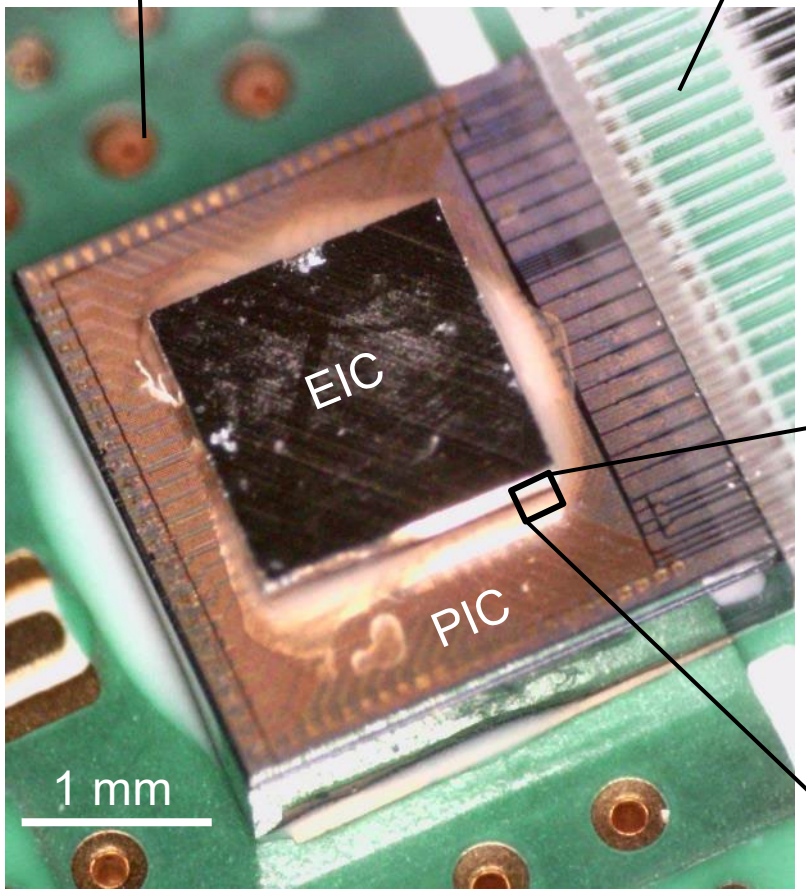
# 3D EIC/PIC Heterogeneous Integration

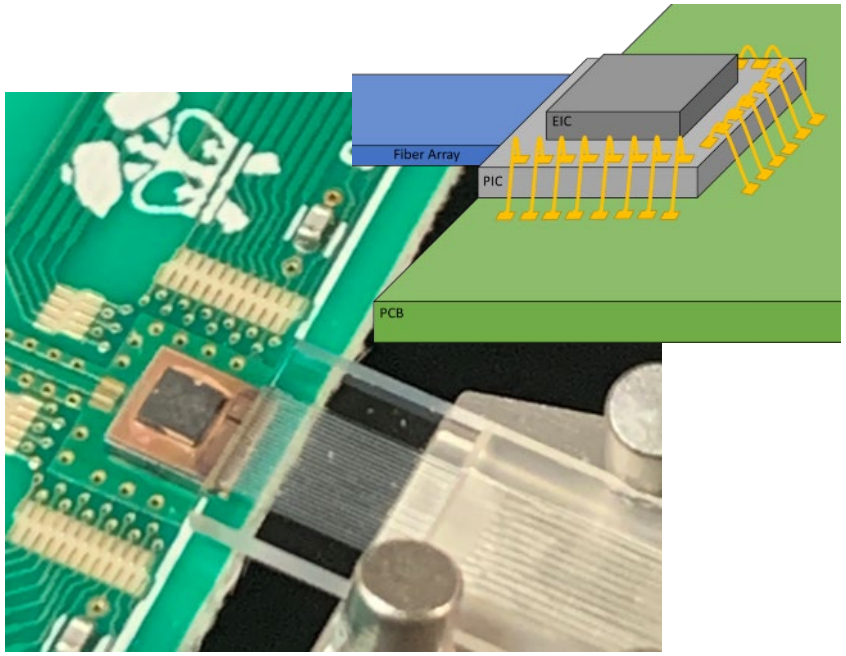
EIC Before Post-Process

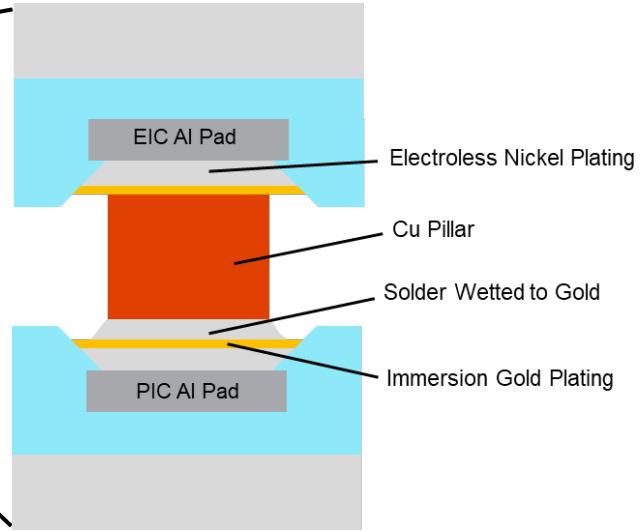



Copper Pillar Bumped EIC



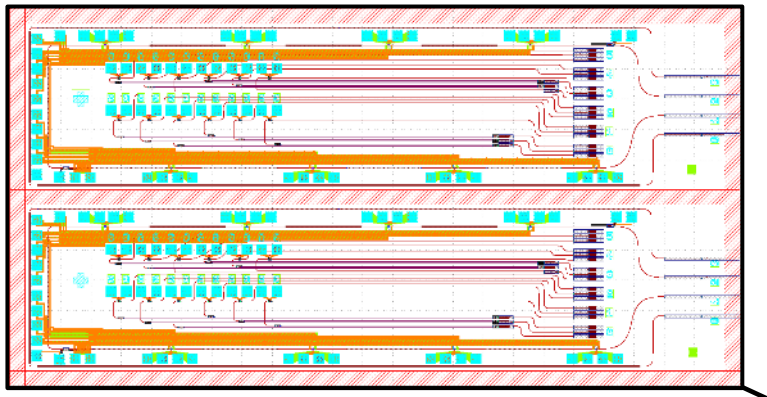



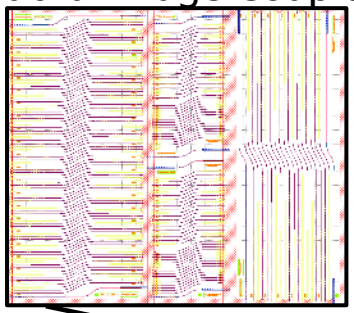



# Full 300 mm Custom Wafer Cedar

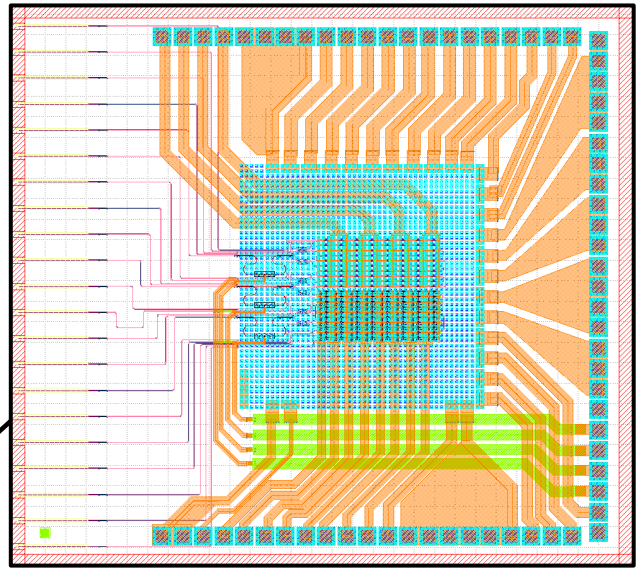
FPGA-packaged WDM Transmitters



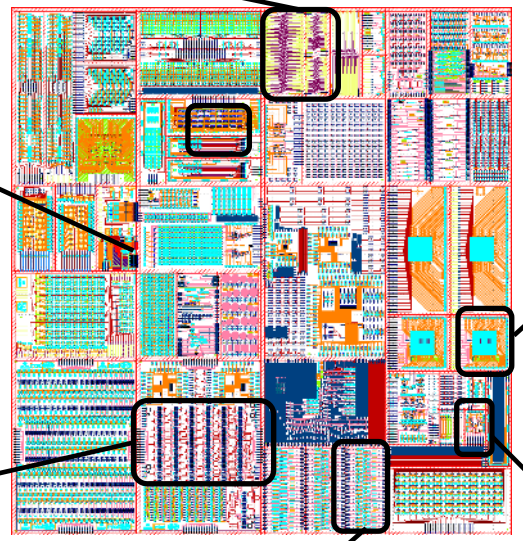
Sub-dB Edge Couplers



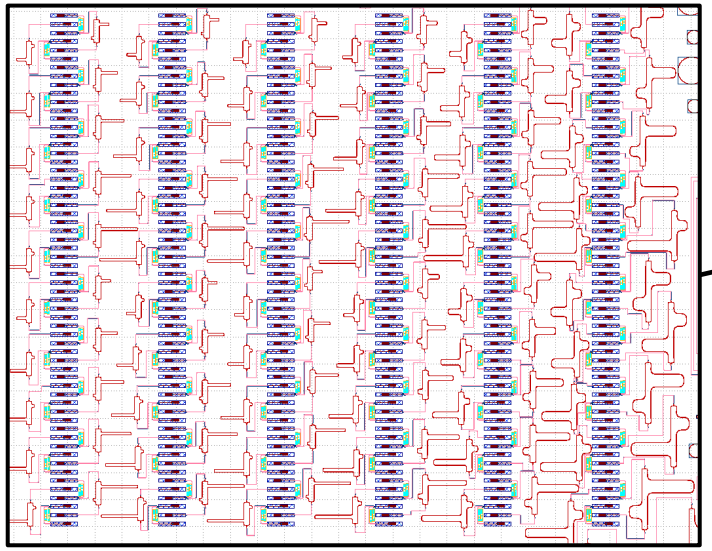
MCM with Custom Modulators



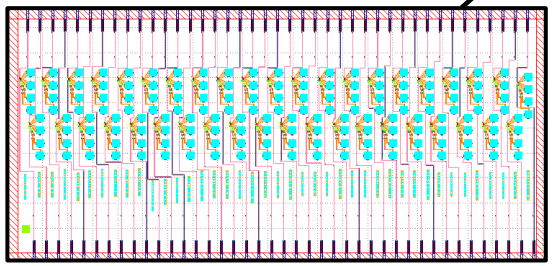
**Cedar**



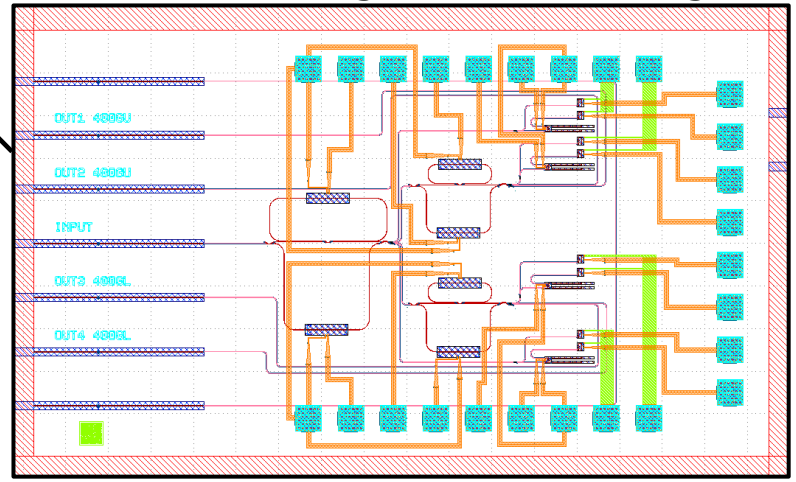
Wafer-scale Quantification of Fabrication Robust Platform Phase Errors



Undercut Modulators



Cascaded RMZI Interleavers with Automated Alignment & Tracking



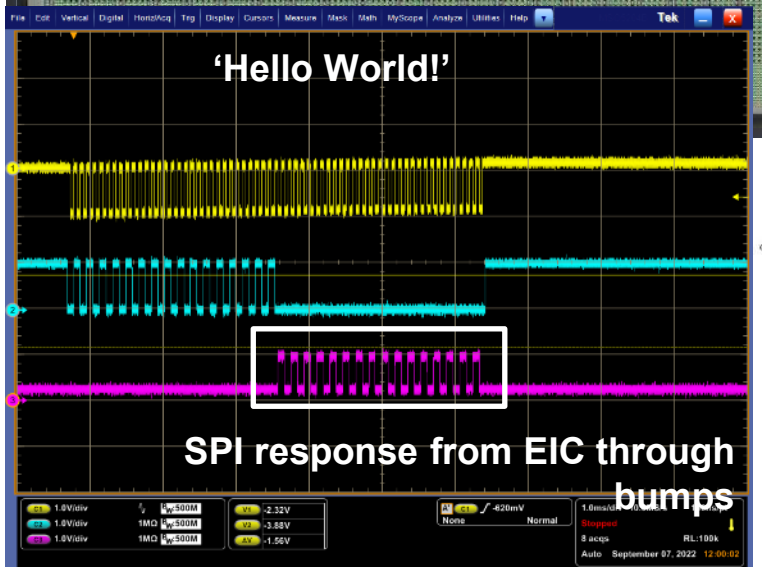
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

# 3D Bumped PIC Wafer + EIC

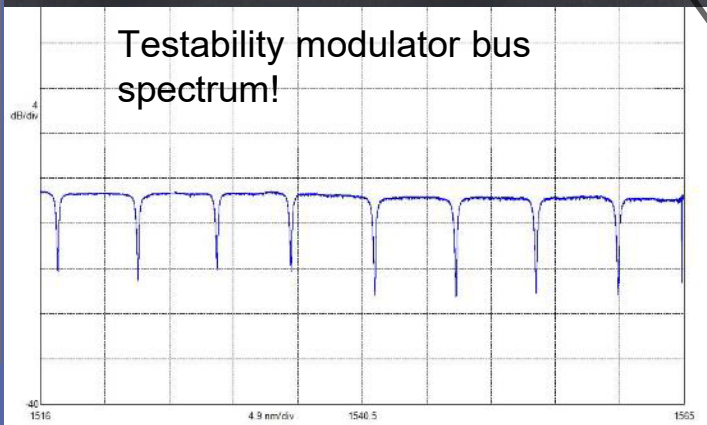
20 buses optically fanned out for testability, yield, and density demonstration

28nm EIC

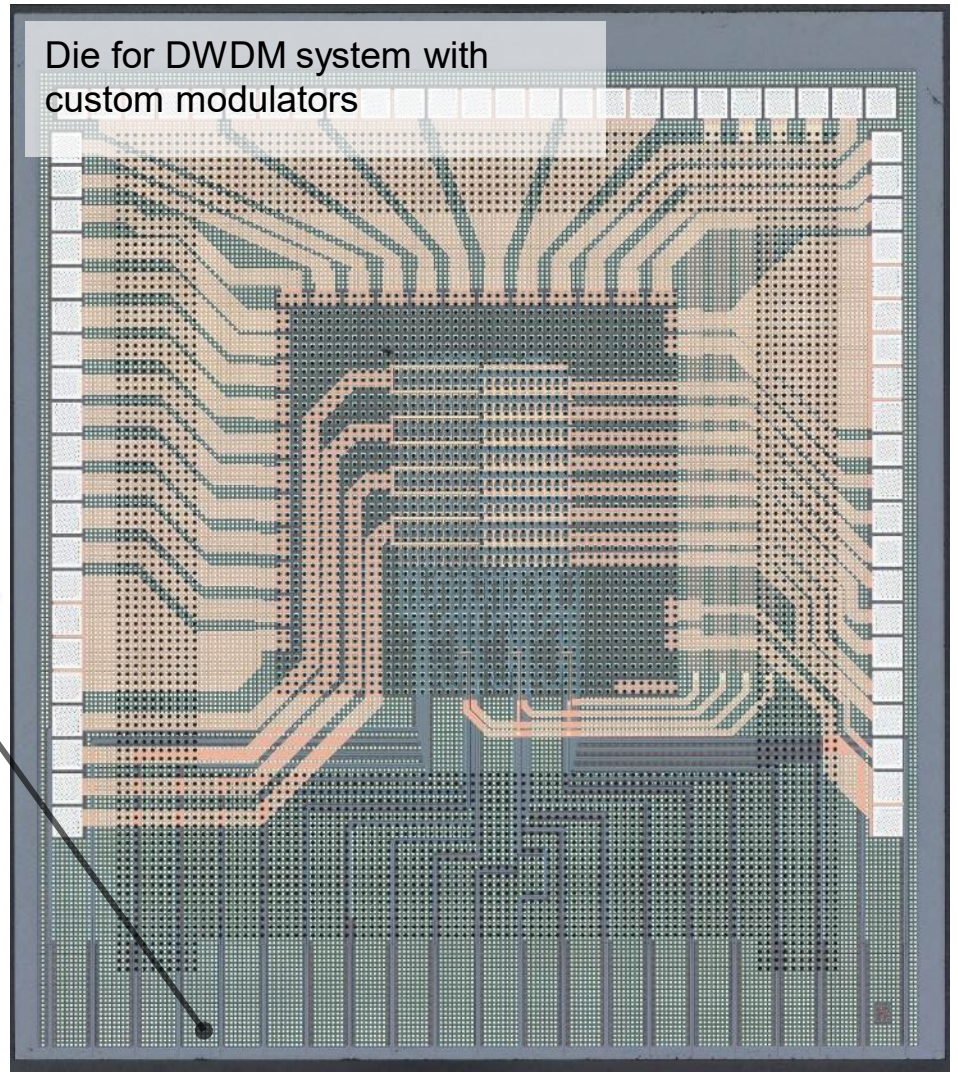
Low loss edge couplers



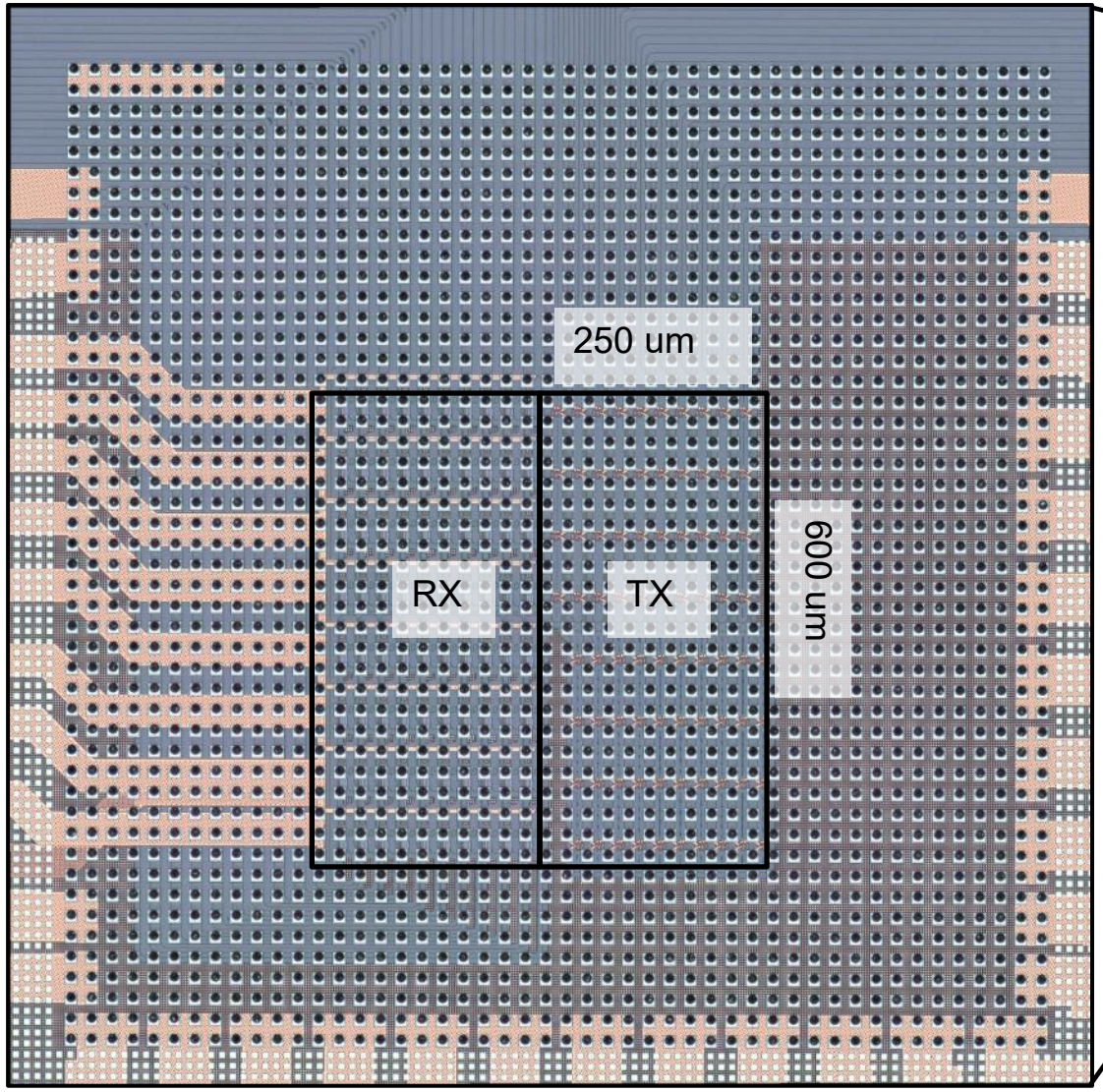

Testability modulator bus spectrum!

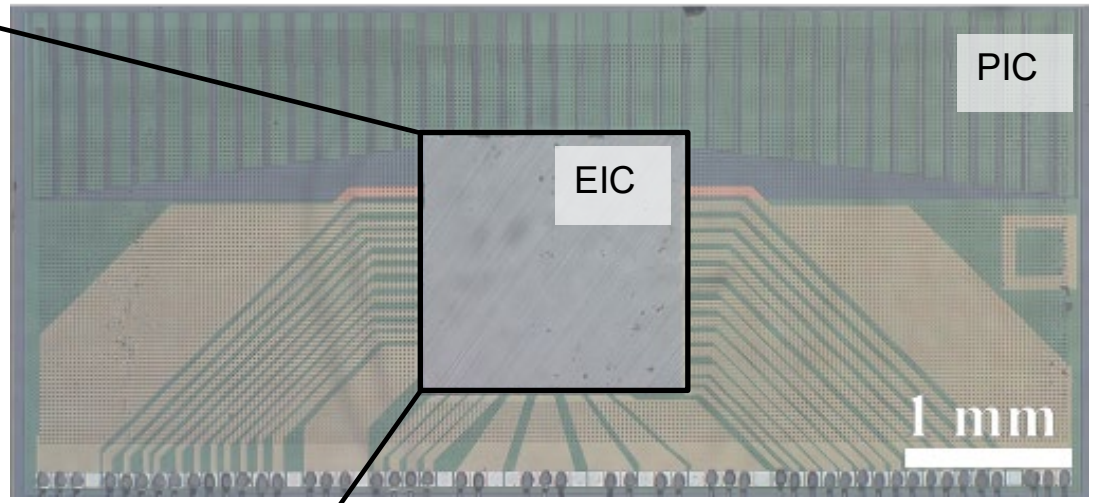


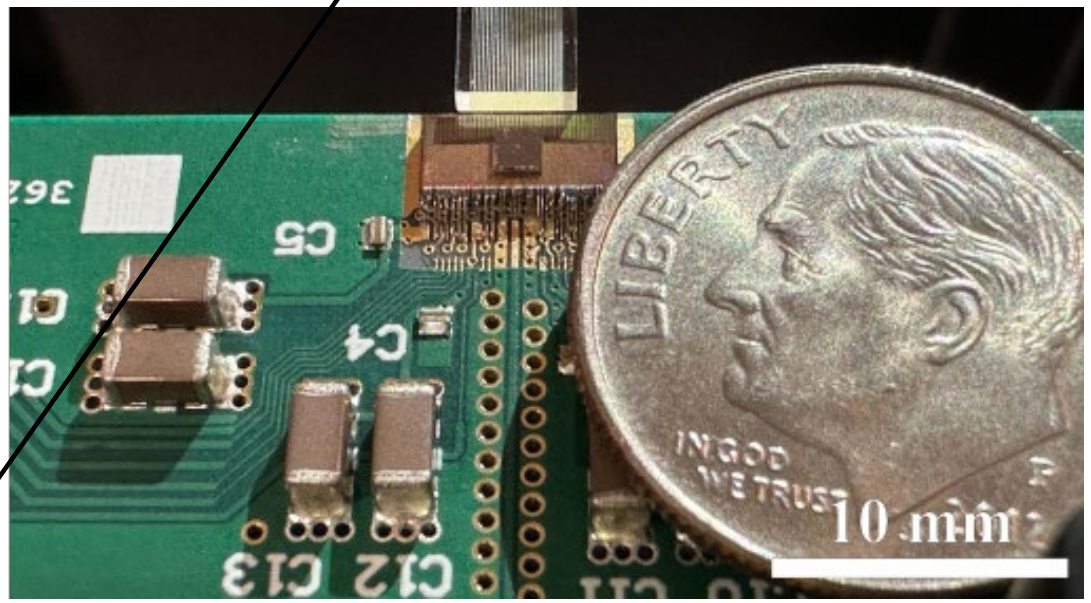
Die for DWDM system with custom modulators



# Fully Assembled High Density 5.3 Tb/s/mm² MCM

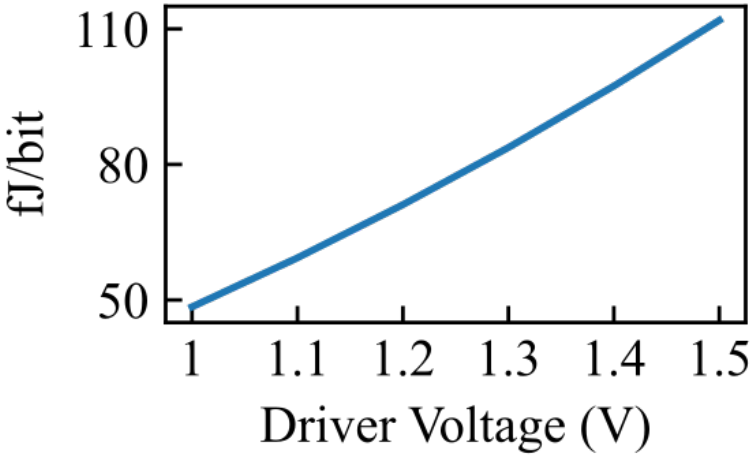


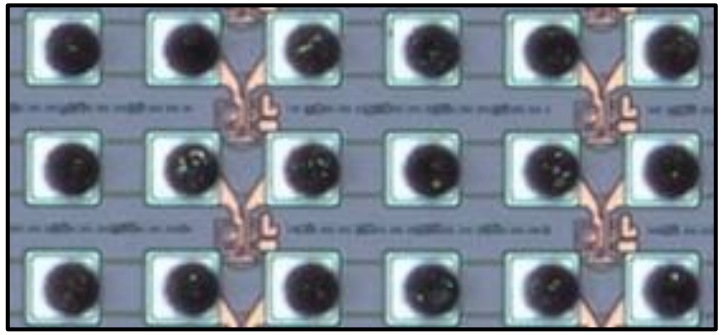



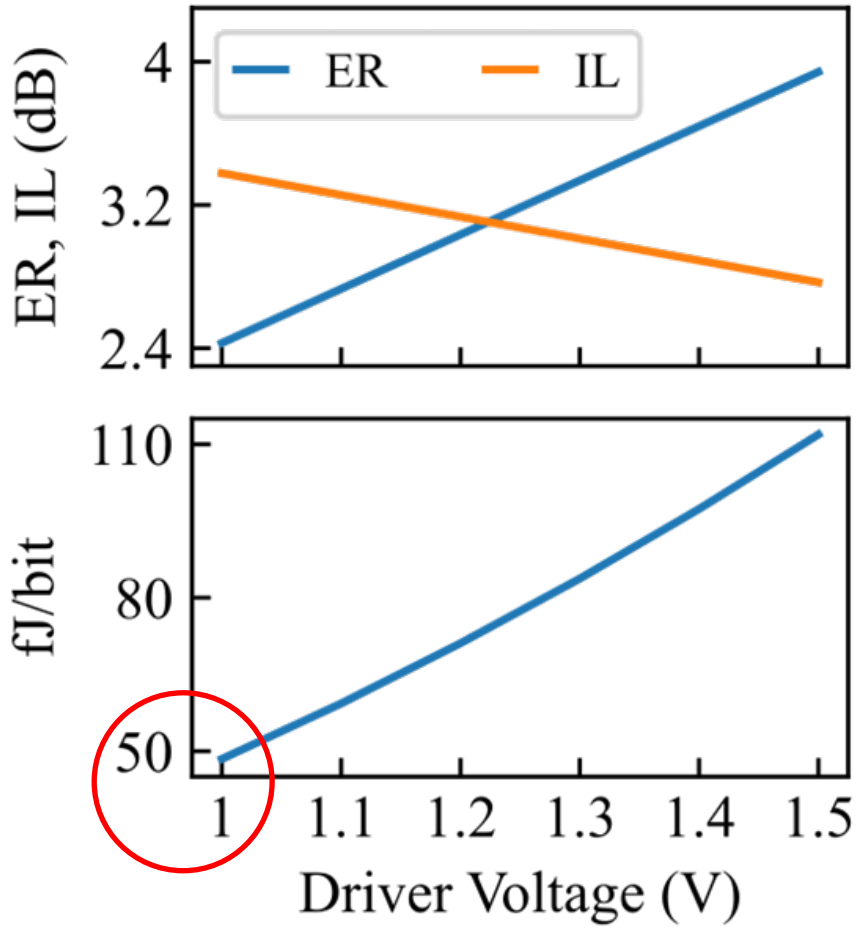
# 800 Gbps at 50 fJ/bit Dense Transmitter Array

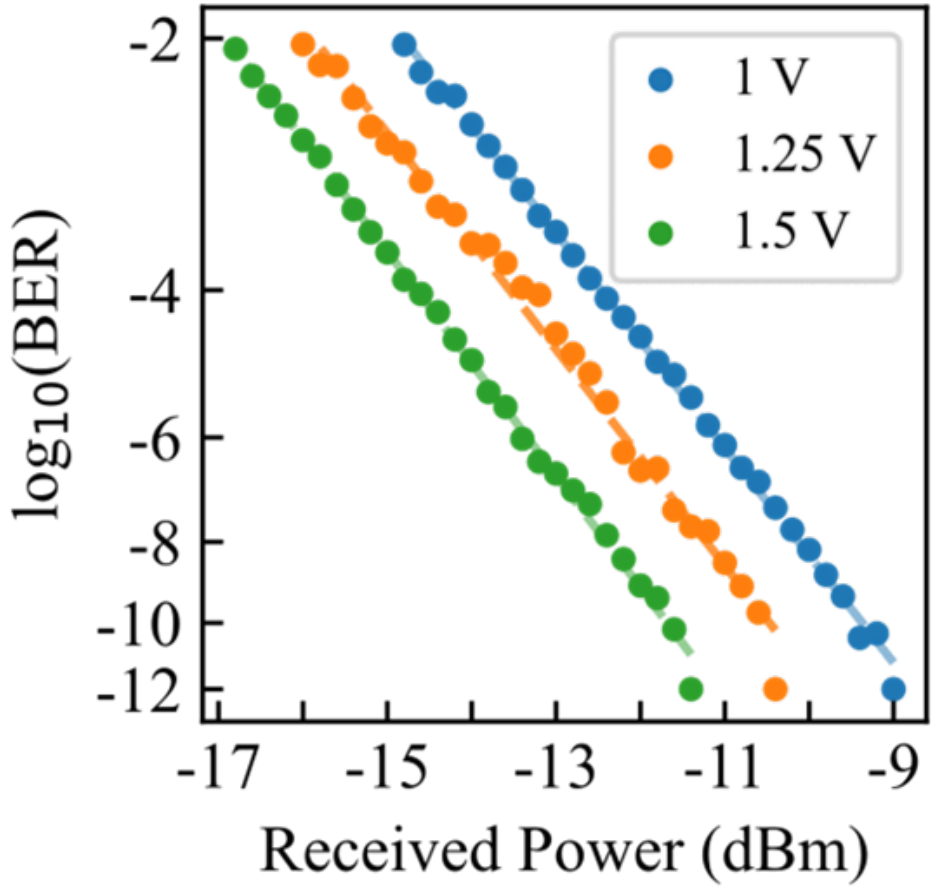

80 Transmitters at 10 Gbps/modulator for **50 fJ/bit**



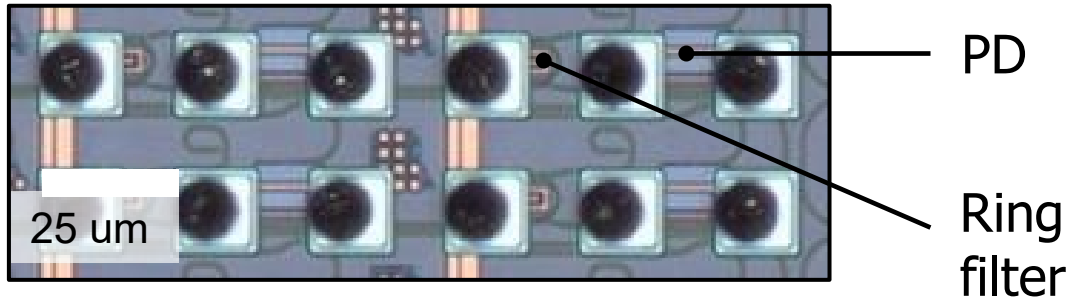
# Realizing 50 fJ/bit Transmitter; BER = 10E-12 and 1Vpp



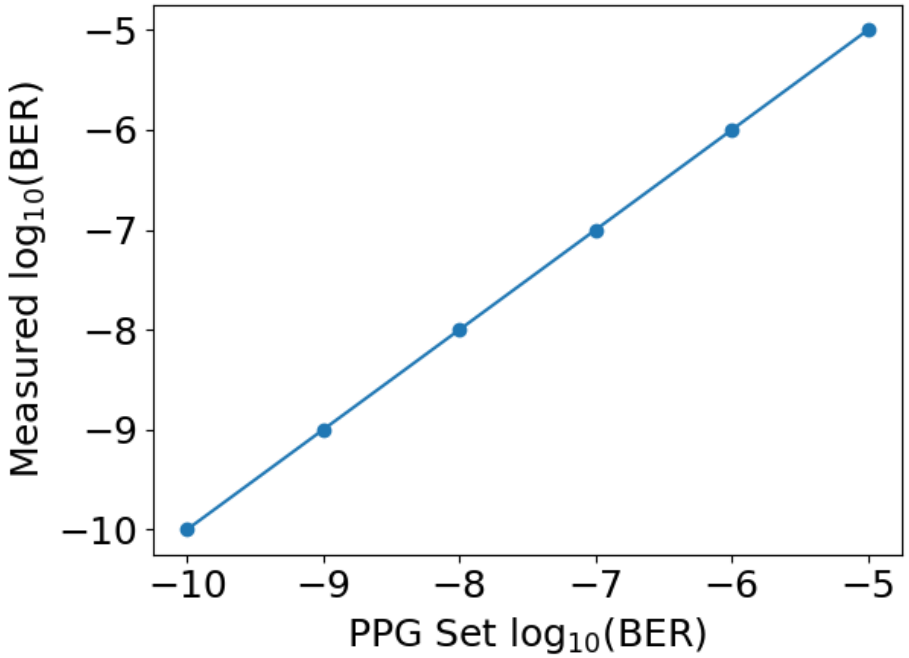




S. Daudlin, S. Lee, D. Khilwani, C. Ou, A. Rizzo, S. Wang, M. Cullen, A. Molnar, and K. Bergman, "Ultra-dense 3D integrated 5.3 Tb/s/mm$^2$ 80 micro-disk modulator transmitter" in OFC 2023, paper M3I.1

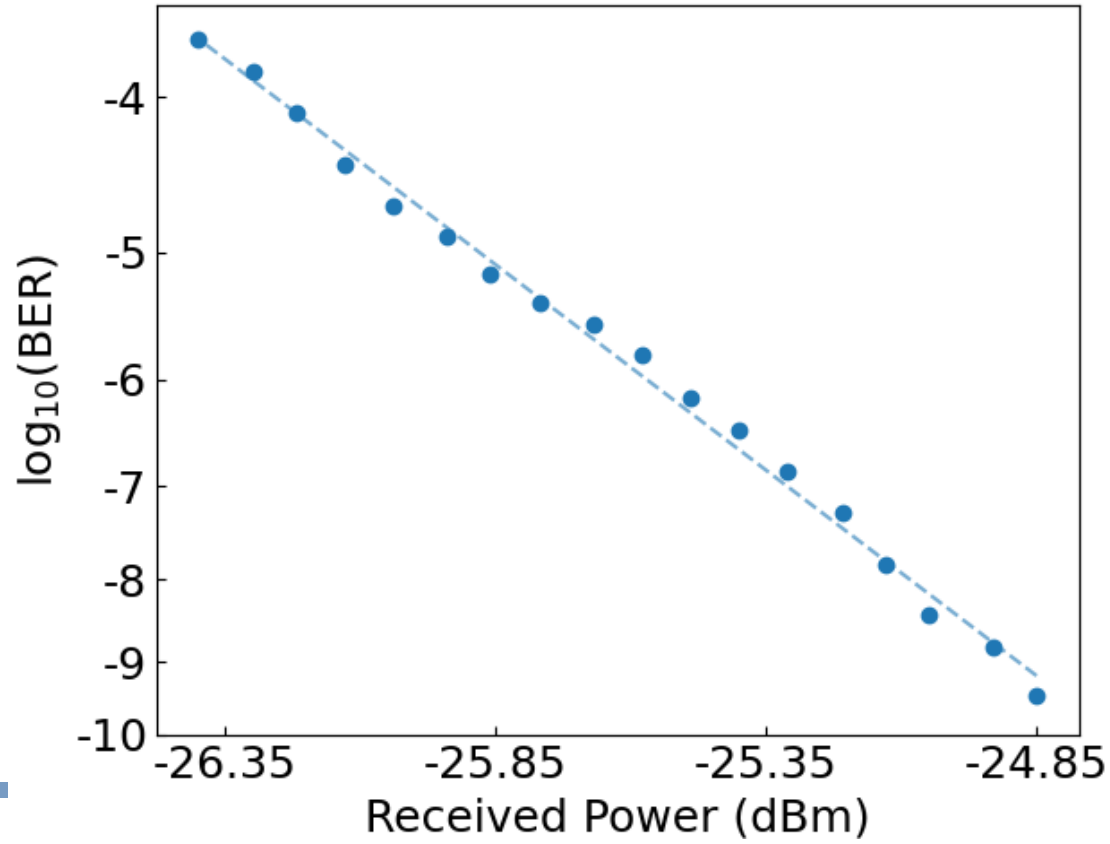
# 68 fJ/bit and -24.85 dBm Sensitivity 800 Gbps Receiver Array




- ✓ On chip error checker accurately detects injected

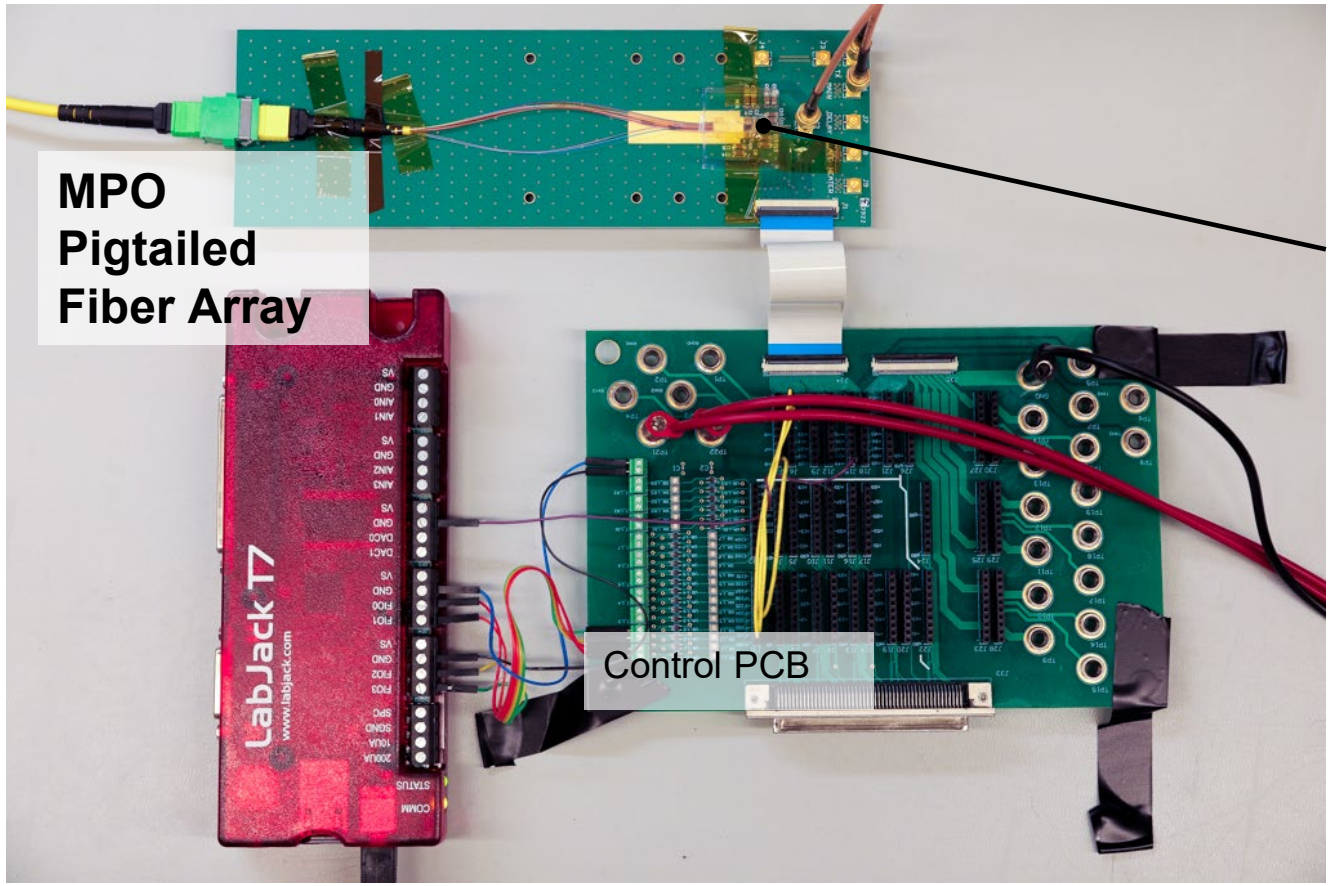


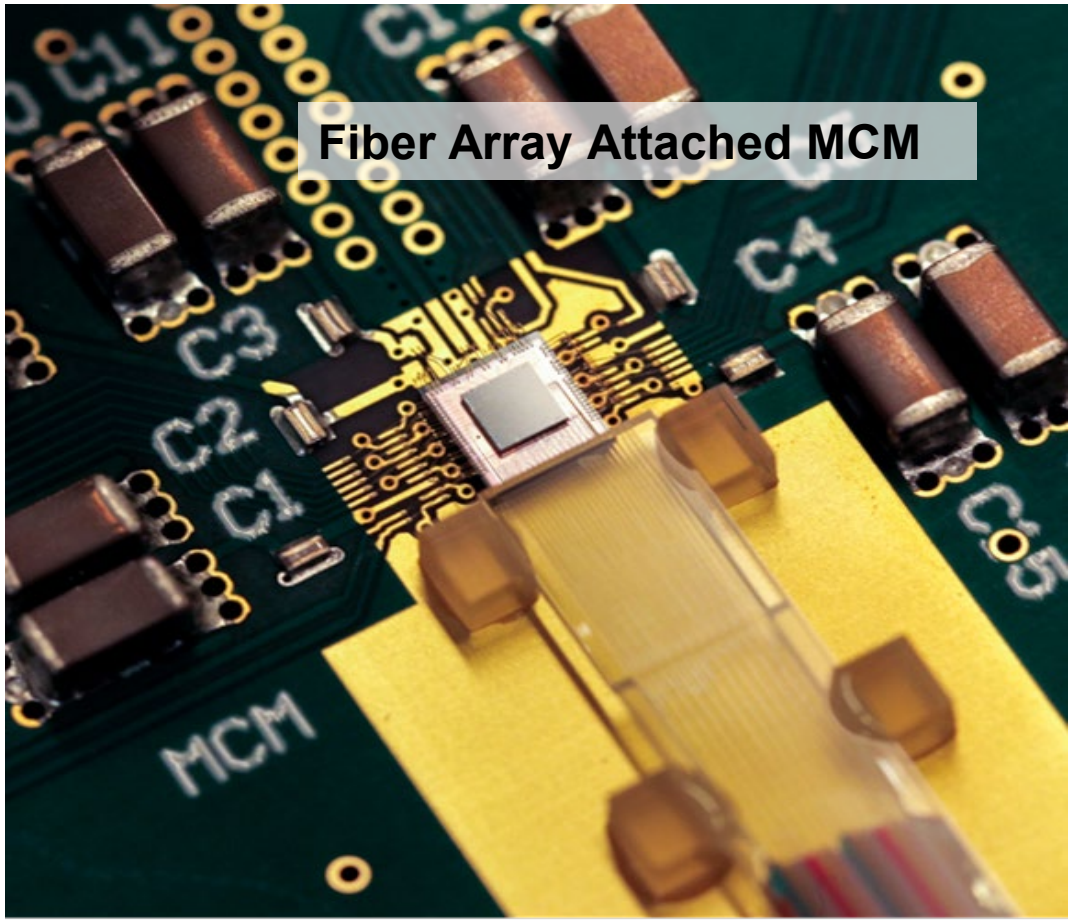

- ✓ Received MZI 10 Gbps transmitted signal at -24.85 dBm with 1e-10 BER
- ✓ 49 fJ/bit static + 19 fJ/bit dynamic power = **68 fJ/bit**



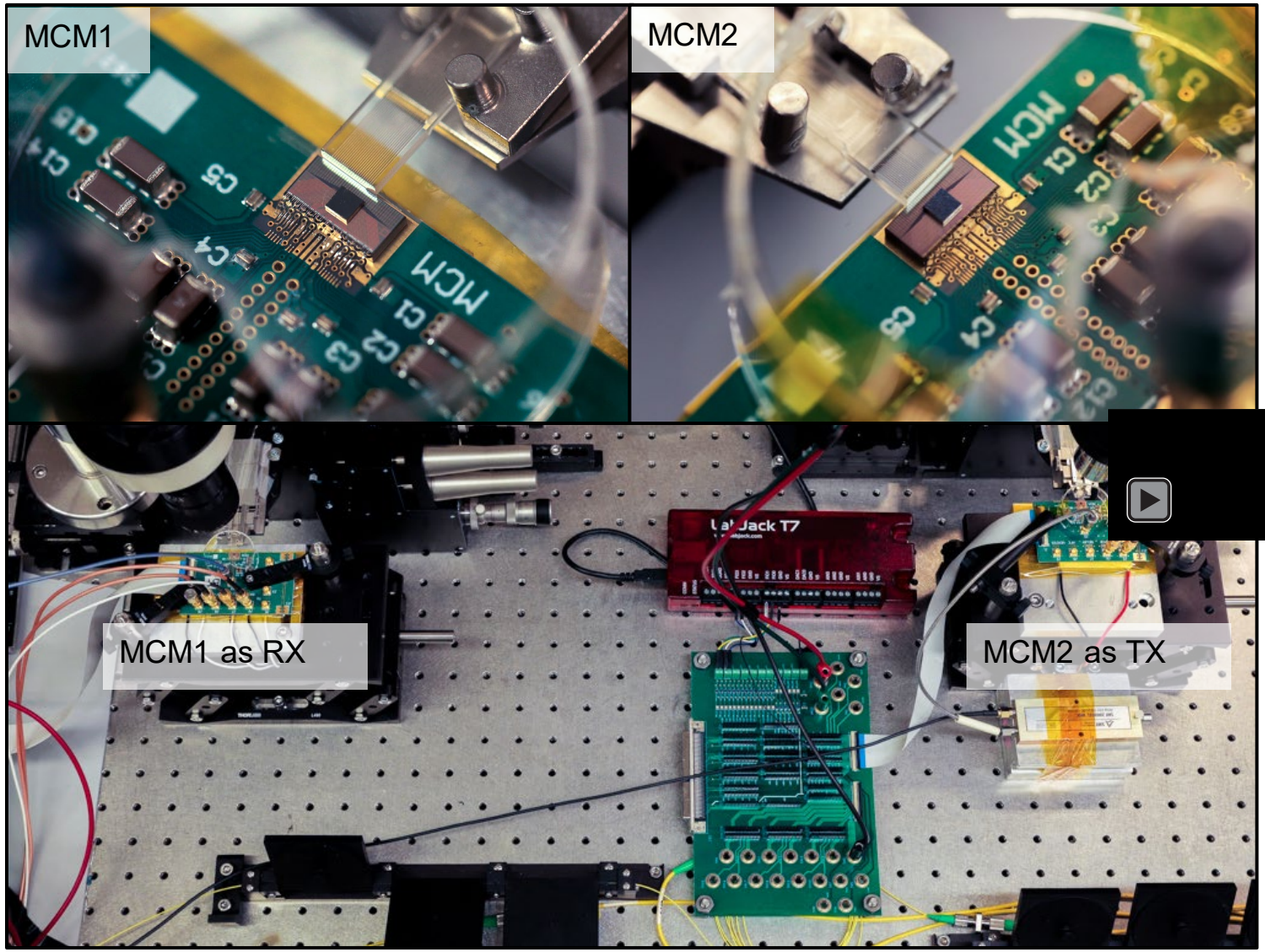
# Fully Packaged MCM with Fiber Array

- ✓ Complete packaging of 3-D integrated MCM with wire-bonding and SMF28 fiber array attach



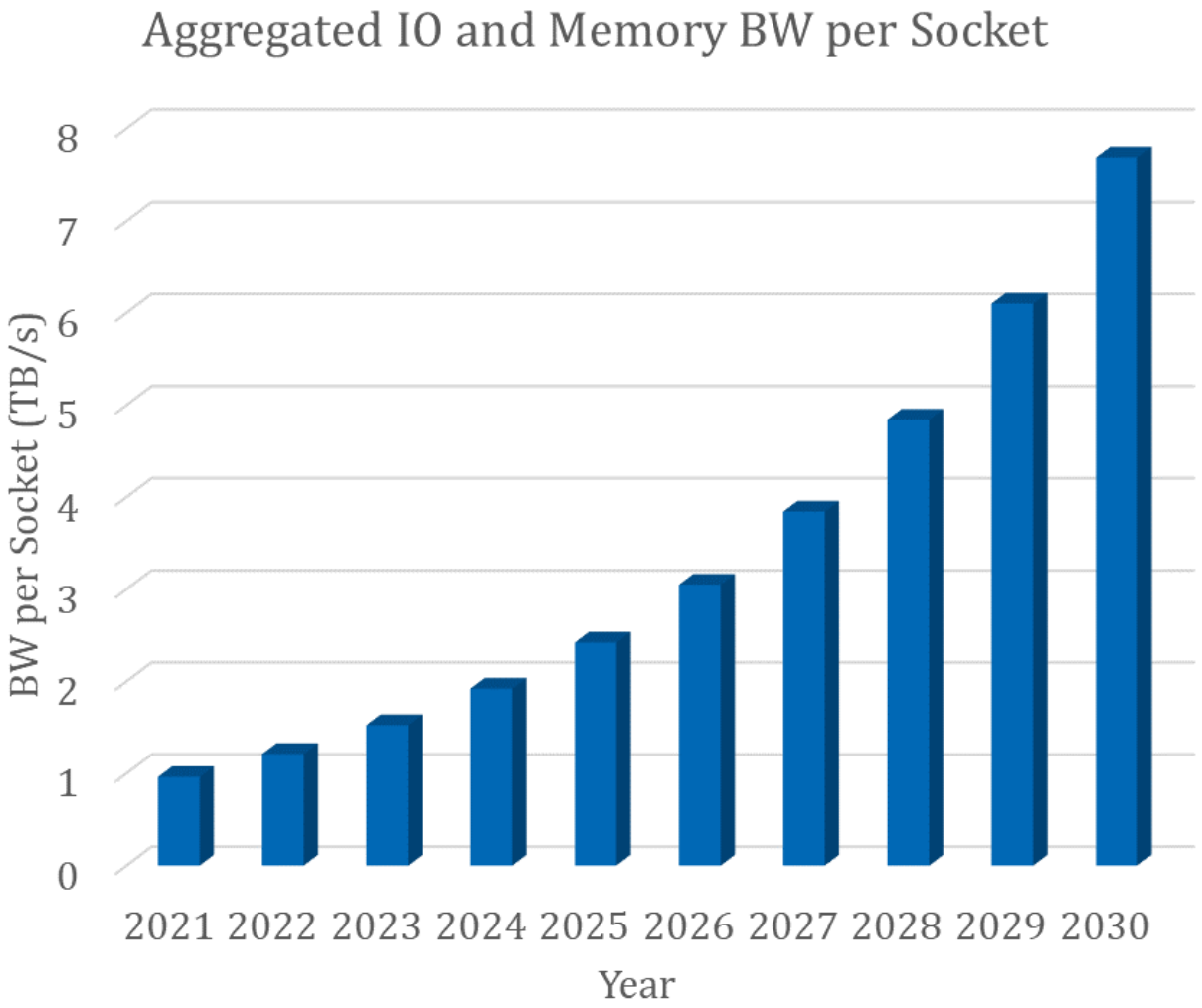



# MCM TX to MCM RX over 100 meters




- ✓ MCM2 transmitting signal to separate MCM1 receiver
- ✓ no amplifiers
- ✓ 100 meters
- ✓ 8 Gbps / channel
- ✓ -6 dBm laser power

# Need for Data Movement is Growing



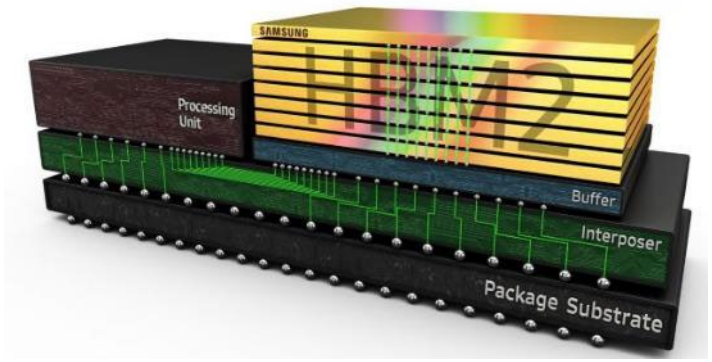

IO + Memory Bandwidth
Scale 2X every 3 years

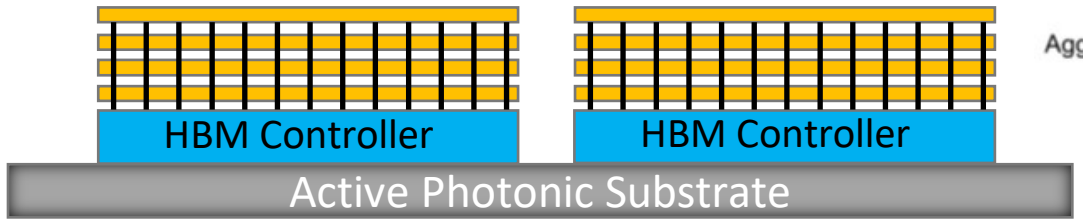
[Fabrizio Petrini, Intel]

# Embedded Photonics – Scaling Ultra-low Energy Memory BW

Samsung Flashbolt HBM †

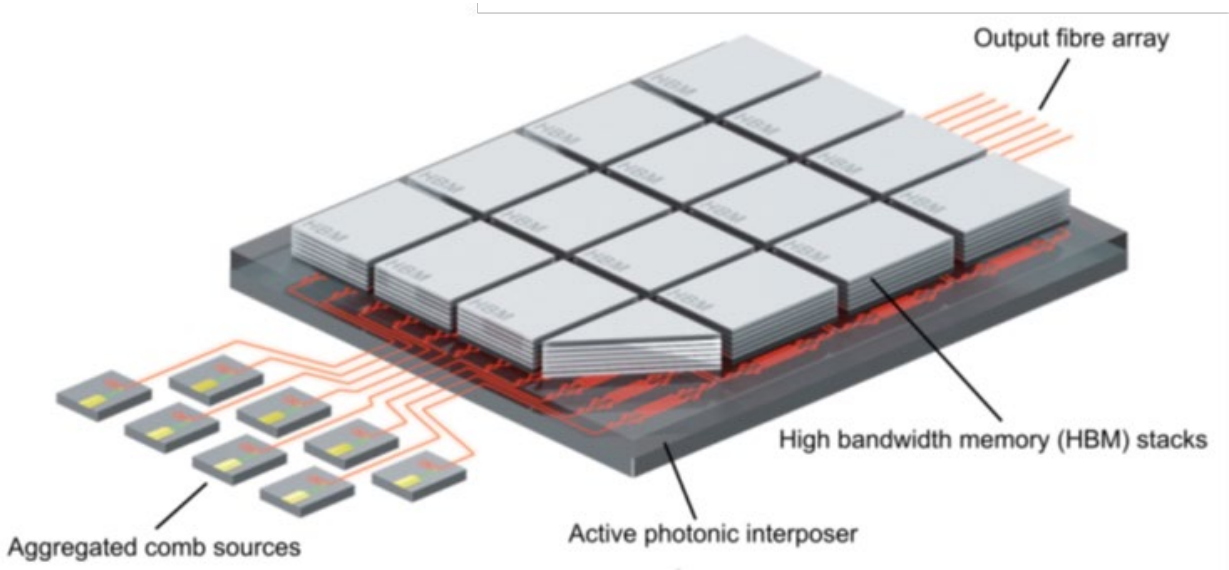
- Capacity 16GB/stack,
- Memory BW ~400GB/s/stack
- Memory BW/capacity ratio: 25x
- 10x11mm = 110mm$^2$




Scaling HBM over full interposer:

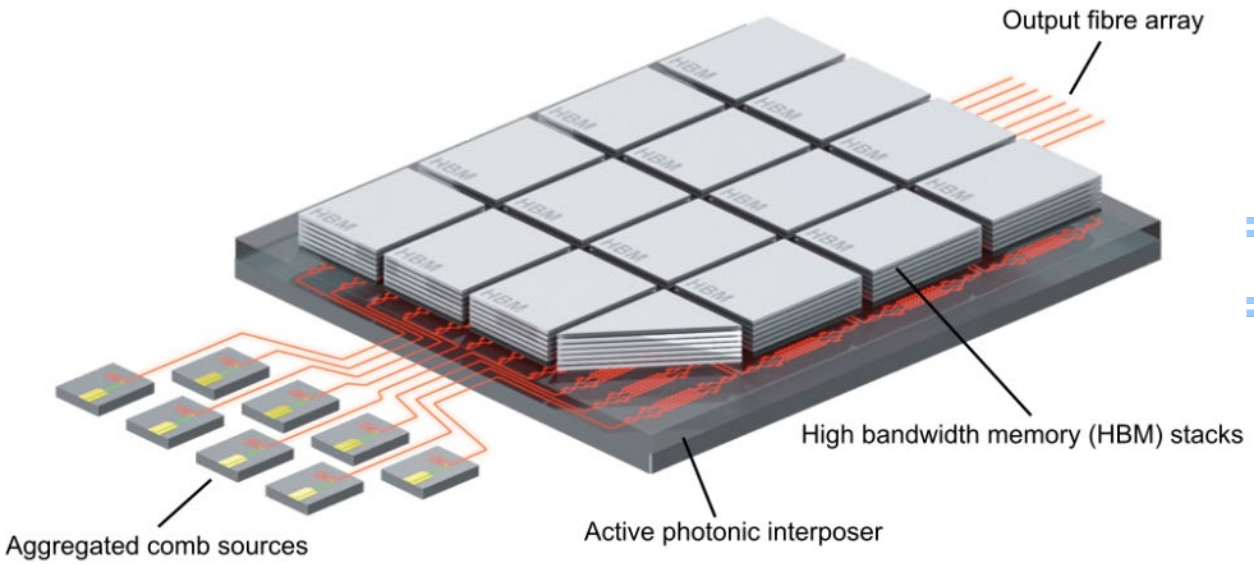
- ~1000mm$^2$ with 9 stacks
- 144GB per package with current HBM
- Using 25x memory BW/capacity ratio: ~4 TB/s

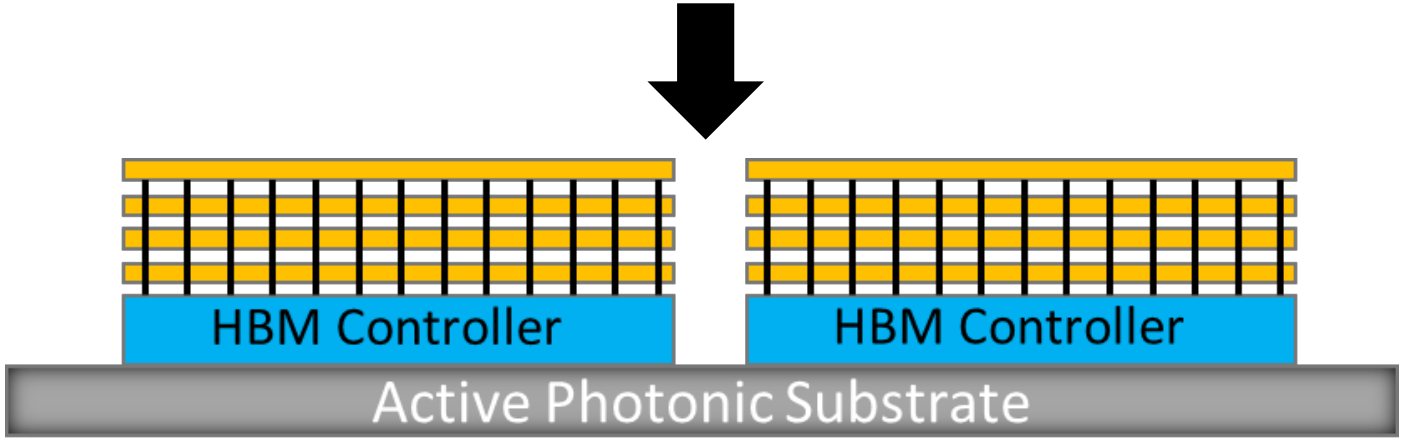


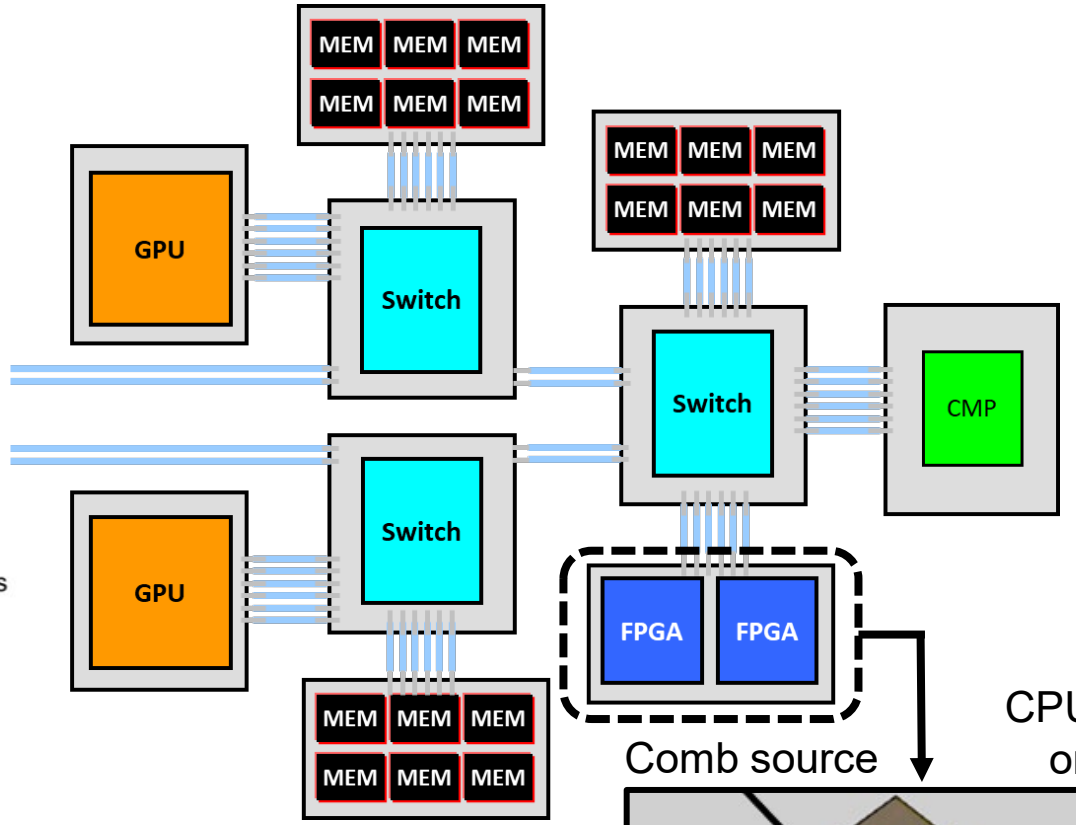

***Optical HBM***

†https://www.samsung.com/semiconductor/dram/hbm-flashbol

# Embedded Photonic Connectivity

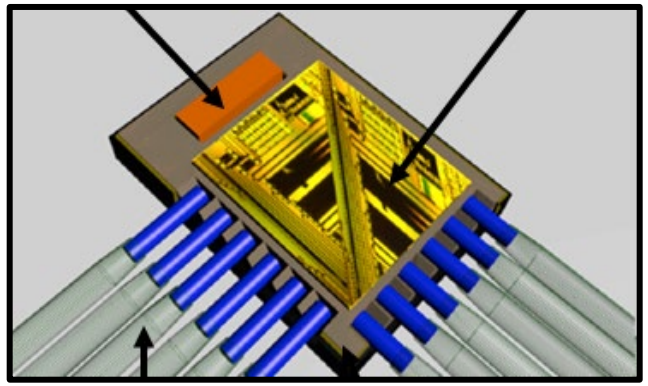



Comb source

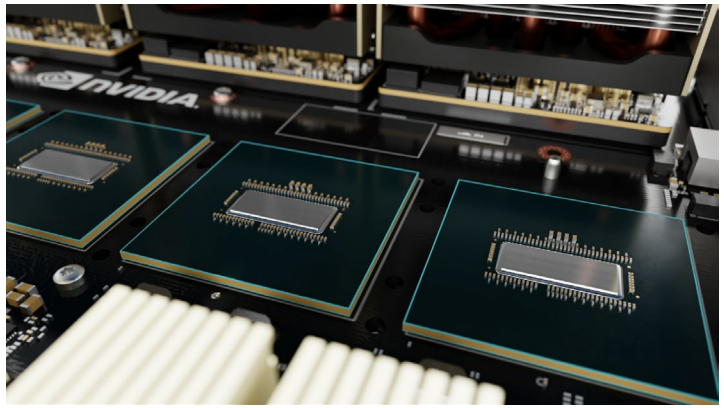
CPUs, GPUs or router

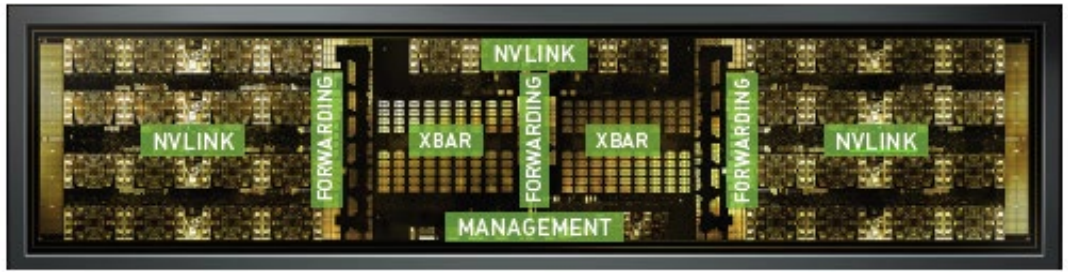


Optical fibers

Photonic die

# Embedded Photonics – Flexible Interconnect Fabric

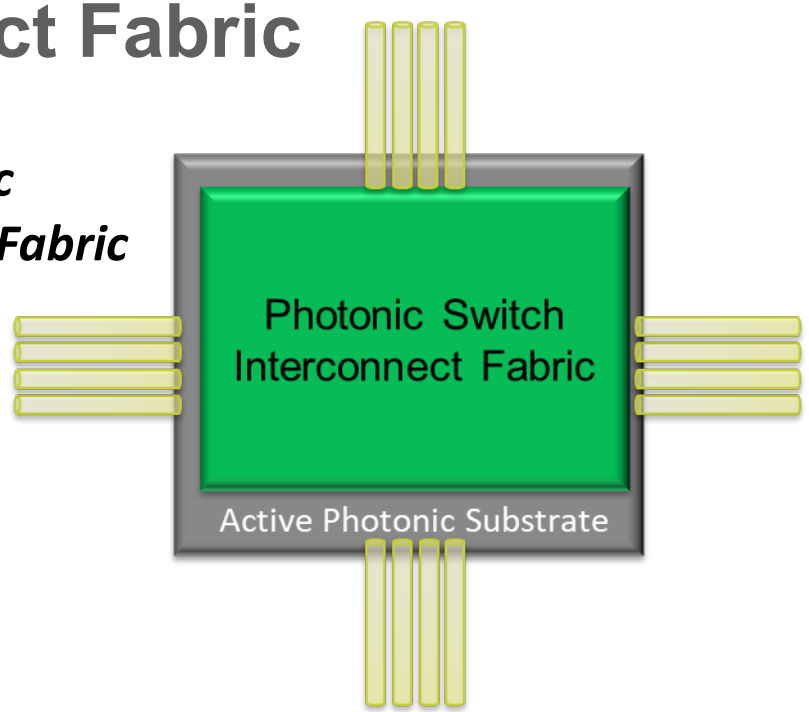



GPU-GPU Interconnect

- NVIDIA A100 DGX – 6 NVSwitches
- 8 GPU DGX system - aggregate BW 4.8TB/s
- Each NVSwitch ~800 GB/s **~ 20 GB/s/mm²**

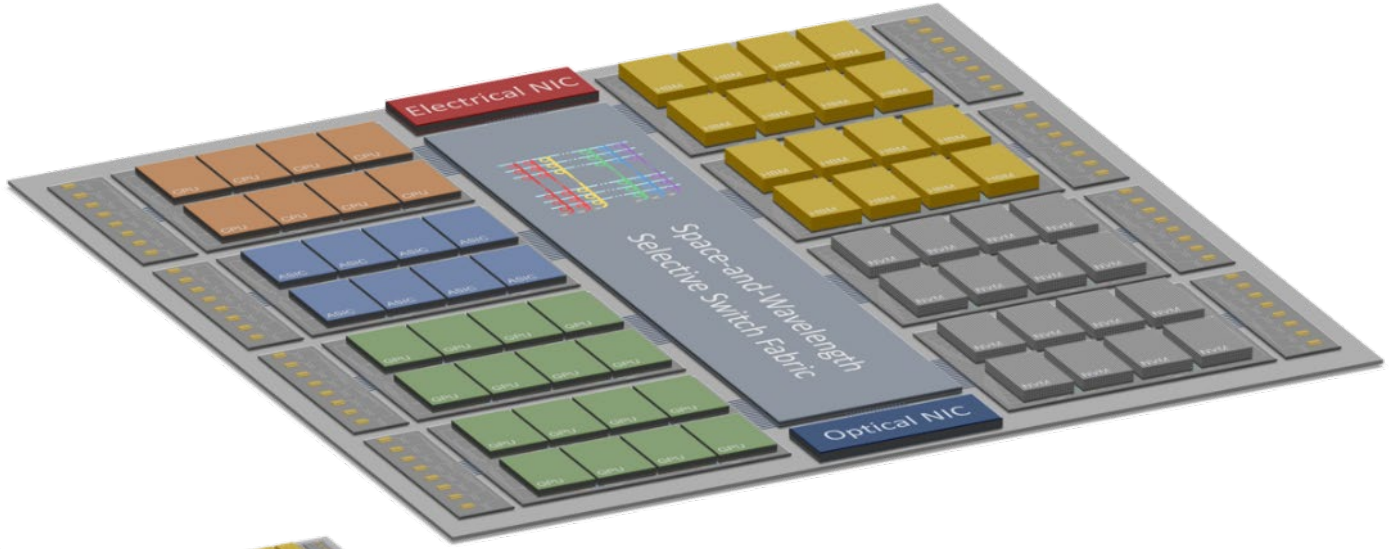
***Photonic Interconnect Fabric***

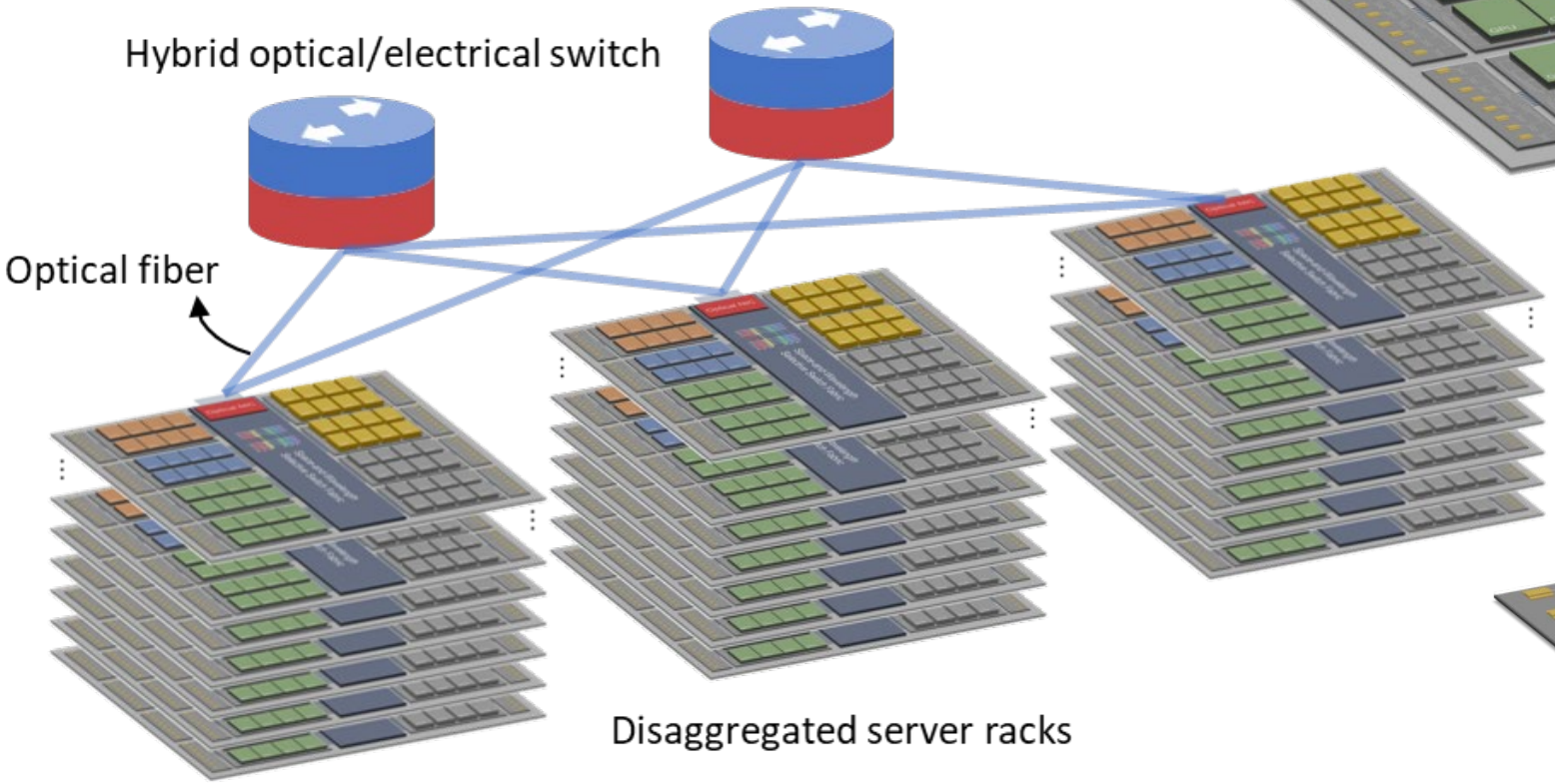



**Photonic Fabric:**

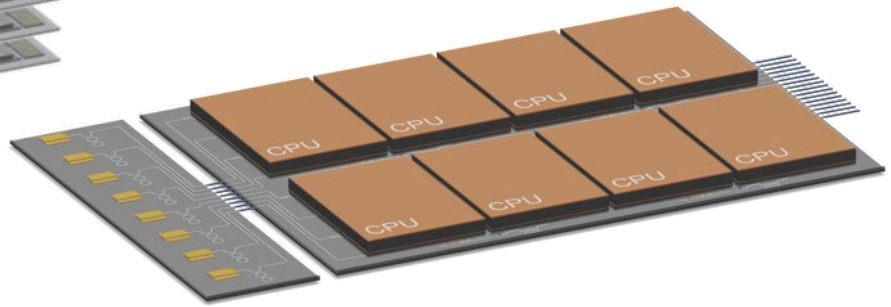
- Target **1.28 TB/s/mm²** in ~1000mm² substrate
- Photonic fabric – **aggregate BW 1280 TB/s**
- 128 ports X 10 TB/s/port
- **Flexible** Spatial/Wavelength/Mode *granularity*
- DARPA LUMOS on-chip gain for scaling
- Optical Multicasting + through on-chip NLO

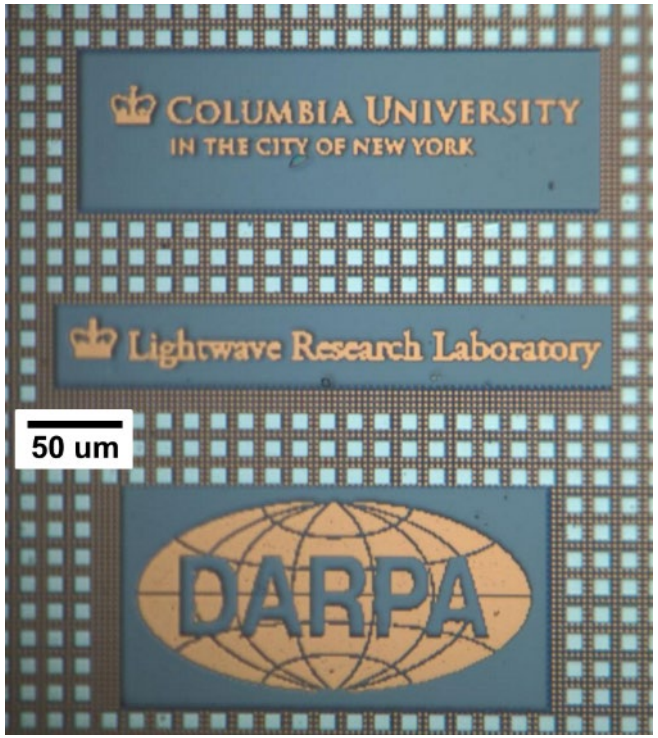
# Scaling Photonic Connectivity System-Wide




**100 Tbps-class optical I/O**




SRC CUbiC DARPA

AFRL THE AIR FORCE RESEARCH LABORATORY LEAD | DISCOVER | DEVELOP | DELIVER

arpa·e CHANGING WHAT'S POSSIBLE

U.S. DEPARTMENT OF ENERGY

DEPARTMENT OF DEFENSE UNITED STATES OF AMERICA

AIM photonics