Realizing Petabit/s IO and sub-pJ/bit System-wide Communication with Silicon Photonics

Keren Bergman
Department of Electrical Engineering
Columbia University, New York, NY
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
Current System Architectures

- GPU to GPU and HBM *intra-group* ~1000 GB/s aggregate bidirectional bandwidth (fat tree).
- *Inter-group* communication relies on ~400 Gb/s links; much slower than the *intra-group* fabric.
- Communication time $\rightarrow 10 \times$ Computation time for DDL workloads trained on > 256 GPUs

[1] NVIDIA DGX SuperPOD: Scalable Infrastructure for AI Leadership
Challenges Moving Data Off-Chip

**GPU-Memory Bandwidth**
- AMD MI300X: 5.2 TB/s
- Nvidia DGX H100: 3.35 TB/s
- Intel Data Center GPU Max: 3.28 TB/s

*Images sources: AMD, Intel, and Nvidia.*
Challenges Moving Data Off-Chip

**GPU-Memory Bandwidth**
- AMD MI300X: 5.2 TB/s
- Nvidia DGX H100: 3.35 TB/s
- Intel Data Center GPU Max: 3.28 TB/s

**GPU-GPU Bandwidth**
- AMD Infinity Fabric: 800 GB/s
- Nvidia NVLink & NVSwitch: 900 GB/s
- Intel Xe Link: 720 GB/s

Images sources: AMD, Intel, and Nvidia.
Challenges Moving Data Off-Chip

**GPU-Memory Bandwidth**
- AMD MI300X: 5.2 TB/s
- Nvidia DGX H100: 3.35 TB/s
- Intel Data Center GPU Max: 3.28 TB/s

**GPU-GPU Bandwidth**
- AMD Infinity Fabric: 800 GB/s
- Nvidia NVLink & NVSwitch: 900 GB/s
- Intel Xe Link: 720 GB/s

**Off-Socket Link Bandwidth**
- InfiniBand: 400 Gb/s
- Projected 800 Gb/s near future

Images sources: AMD, Intel, and Nvidia.
Challenges Moving Data Off-Chip

**On-Chip**
- **GPU-Memory Bandwidth**
  - AMD MI300X: 5.2 TB/s
  - Nvidia DGX H100: 3.35 TB/s
  - Intel Data Center GPU Max: 3.28 TB/s

**In-Socket**
- **GPU-GPU Bandwidth**
  - AMD Infinity Fabric: 800 GB/s
  - Nvidia NVLink & NVSwitch: 900 GB/s
  - Intel Xe Link: 720 GB/s

**Off-Socket**
- **Off-Socket Link Bandwidth**
  - InfiniBand: 400 Gb/s
  - Projected 800 Gb/s near future

Off-Socket IO BW limit creates **100 X Bandwidth Taper** across system

Images sources: AMD, Intel, and Nvidia.
Bringing Photonics to the Chip

Adapted from Gordon Keeler, DARPA
Bringing Photonics to the Chip

2.5D Integration

Pros:
• Better density than 2D
• Balanced scalability & flexibility
• Thermal isolation

Cons:
• Parasitics from doubled bump interfaces and traces
• Still limited BW density
• Added complexity from interposer design

~400 Gbps/mm
~10 pJ/b
Bringing Photonics to the Chip

Monolithic Integration

Pros:
- Minimal parasitics
- Simplified packaging
- Thermal dissipation

Cons:
- Bandwidth density limited by electronics
- Outdated technology nodes limit power, scaling

Source: Ayer Labs

2.5D Integration

<table>
<thead>
<tr>
<th></th>
<th>~400 Gbps/mm</th>
<th>~10 pJ/b</th>
</tr>
</thead>
</table>

Monolithic Integration

|               | ~200 Gbps/mm | ~5 pJ/b  |
Bringing Photonics to the Chip

3D Integration

Advantages:
- Best shoreline & area bandwidth density
- Massive wavelength scalability
- Benefits from advanced CMOS technology nodes

Challenges:
- Packaging yield
- Thermal management

<table>
<thead>
<tr>
<th>2.5D Integration</th>
<th>Monolithic Integration</th>
<th>3D Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>~400 Gbps/mm</td>
<td>~200 Gbps/mm</td>
<td>Multi-Tbps/mm</td>
</tr>
<tr>
<td>~10 pJ/b</td>
<td>~5 pJ/b</td>
<td>Sub-pJ/b</td>
</tr>
</tbody>
</table>
Silicon Photonics Fabrication

- Low Loss Chip Coupling
- 300 mm SOI Wafers
- High Speed Modulators
- High Performance Passives (splitters, filters, polarization control)
- Ge Detectors
- Wavelength interleaving
Photonics = Massive Parallelism in the Wavelength Domain

Frequency Combs: Multi-Tb/s per Single Link

Anthony Rizzo, Asher Novick, Vignesh Gopal, Bok Young Kim, Xingchen Ji, Stuart Daudlin, Yoshitomo Okawachi, Qixiang Cheng, Michal Lipson, Alexander L. Gaeta & Keren Bergman, “Massively scalable Kerr comb-driven silicon photonic link” Nat. Photon. (June, 2023)
Approach to reaching multi-Tbps IO and sub-pJ/b

**Key Technical Innovations:**

- Embrace extreme parallelism:
  - Ultra-dense channels generated by > 100 wavelengths (DWDM) 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
  - Wafer scale undercut for increased efficiency
Full 300 mm Custom Wafer Cedar

FPGA-packaged WDM Transmitters
Wafer-scale Quantification of Fabrication
Robust Platform Phase Errors

Sub-dB Edge Couplers
Undercut Modulators

Cedar

MCM with Custom Modulators
Cascaded RMZI Interleavers with Automated Alignment & Tracking
Full 300 mm Custom Wafer Oak Tapeout

- Polarization Rotator Splitters
- Non-packaged Link Test Structures
- Fully packaged 4x4x64 Links
- Package-able Resonator Arrays for Tx and Rx
- Novel Spatial-Spectral Switch Structures
- Oak high speed custom disks

10Gbps
16Gbps
20Gbps
25Gbps
Fully Packaged MCM with Fiber Array

- Complete packaging of 3-D integrated MCM with wire-bonding and SMF28 fiber array attach
System Wide Seamless Connectivity with Embedded Photonics

Stojanovic, Wu

4 × 4 × 4λ switch

CPUs, GPUs or router

Optical fibers

Photonic die

Active Photonic Substrate

HBM Controller

HBM Controller

Output fibre array
High bandwidth memory (HBM) stacks
Aggregated comb sources
Active photonic interposer
ONIC Development – FPGA Programmable Photonics Network Interface

PIC – 2 x 16-Channel Transceivers

PCB – RF and Low-Speed Routing

Fibre

PIC

Interposer

EIC

ONIC PCB

FMC

FMC

FMC

PCI Express Gen4 Development Platform PCB with FPGA

HTG 930 – PCIE to Host Server

Receiver BER $< 10^{-12}$ up to 25Gbps/λ

First optically-packaged ONIC, connected to HTG 930
Enabling Adaptable, Disaggregated Architectures

System Scalability with Photonic Connectivity

100 Tbps-class optical I/O

HBM Memory Disaggregation

A. Rizzo et al., Nature Photonics, 2023
Explosive Growth in Data Communication Demands

Cloud Connectivity Challenges:
- Orders of magnitude gap between on-chip/off-chip BW
- Strong distance-dependent communication energy
- Scalability limited by energy and bandwidth tapering
- Massive heterogeneity – compute/memory/accelerator

Edge Connectivity Challenges:
- Driving mm-Wave capacity to meet data demand with robustness, reliability, mobility, and low cost
- Massive densification, power, loss, thermal cooling
- Long-range links - back-haul, long range front-haul, airborne links - limited by output power

System Connectivity Challenges:
- Seamless connectivity between edge and cloud for optimized cross-layer performance
- Reconfigurable, adaptable connectivity to accelerate heterogeneous applications
- Secure and resilient connectivity across edge and cloud
$35M 5-Year Program
Launched Jan 2023

• DARPA (JUMP 2.0)
• 15 companies
• NSF (REU)