PINE Phase 2: Deeply Disaggregated Computing Systems with Embedded Photonics

Keren Bergman

Lightwave Research Lab
Columbia University, New York, NY, USA
High Performance Systems: Trends and Challenges

- **Fugaku (Fujitsu and RIKEN)**
  - Most powerful supercomputer* (June, 2020)
  - Performance: 415.5 PetaFLOPS
    - 80% of peak Linpack
    - 2.8X over Summit
  - Power consumption: 28.33MW
    - Efficiency: 14.7 GFLOPs/Watt (#9 Green 500)
  - 158k Nodes (432 racks) with:
    - ARM A64FX 2.6TFLOP/node
    - 3D stacked memory 4x 8GB HBM 1024 GB/s

- **6D Torus Tofu-D**
  Each node 10 ports and 400Gb/s BW
  Aggregate network BW 6.5PB (52Pb/s)

* at executing High Performance Linpack (top500.org results)

Fugaku performance on HPCG = 13PF
This is 2.8% of the Peak Compute
**Energy Optimized Links**

- **800G aggregate bandwidth per link and 2.2pJ/bit**

**Optically Connected Multi-chip Modules**

- Multi-chip module 2.5D and 3D assembly developed with high-density active interposer **3.2 Tbs/mm**.

**Adaptive Bandwidth Steering**

- PINE system architecture with flexibly assembled nodes for bandwidth steering.

**Deeply Disaggregated Architecture**

- PINE Flexible Fat Tree architecture: efficient resource utilization, tapering, **reduced power-per-transaction**

- Average network latency reduces by 87%
- Average throughput (transactions/sec) improve 4.3x
Diversity of Memory Requirements

About 15% of NERSC workload uses more than 75% of the available memory per node. But 50% use LESS than 15% of memory spike at 100% - these are the jobs that reported memory use that is physically impossible (more than installed memory).

John Shalf, George Michelogiannakis, Brian Austin, Taylor Groves, Manya Gorbadi, Larry Dennison, Tom Gray, Yiwen Shen, Min Yee Teh, Madeleine Glick, and Keren Bergman, “Photonic Memory Disaggregation in Datacenters,” Paper PsW1F.5, OSA Advanced Photonics Congress, July 2020.
Flexible Photonic Interconnected Resources

Embedded Photonics into Heterogeneous Compute/Memory

Approach: network of resources...rather than a network of servers
GPU-Memory

- Each NVIDIA A100 GPU
- 6 HBM stacks per GPU
- 40 GB HBM / GPU
- 1.6 TB/s / GPU

High BW Memory constrained:
- How much can fit in close proximity to GPU
- Same interposer, wirelength, footprint
- Power density


https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/
Embedded Photonics – Scaling Memory BW

Samsung Flashbolt HBM†
- Capacity 16GB/stack,
- Memory BW ~400GB/s/stack
- Memory BW/capacity ratio: 25x
- 10x11mm = 110mm²

Scaling HBM over full interposer:
- ~1000mm² with 9 stacks
- 144GB per package with current HBM
- Using 25x memory BW/capacity ratio: ~4 TB/s

†https://www.samsung.com/semiconductor/dram/hbm-flashbolt
Deep Disaggregation: MCM photonic connectivity/switching

Optically Interconnectivity for Deep Disaggregation
MCM can be reconfigured to accelerate different applications

Integrated Photonic Multi-chip Modules (MCM) Technology
Deep Disaggregation can provide >10:1 dynamic bandwidth range

>5X Performance-per-Watt advantage for applications with diverse node resource demands.

Data Analytics Workloads

Photon MCM Connectivity Map

Virtual “Pin” destination for GPU MCM
PINE SiP Energy Optimized Dense WDM Link

- Passive alignment high-density fiber chip-IO with <0.6 dB coupling penalty >80 degrees

- Energy optimized first fully integrated comb generator with 45% conversion efficiency.

- First monolithic QD SOA on silicon – record WPE at 14.2% and 39dB on-chip gain.
Record usable bandwidth > 125 nm; flat-top response and > 10 dB worst case extinction ratio
2.5D and 3D High-Density Multi-Chip Modules

- Fully Packaged 32-channel WDM Transmitter
- 2.5 dB/facet insertion loss, open eye to 16 Gb/s
SiN Comb Chip

Active SiP Cascaded Modulator Chip

- **First** demonstration of Kerr comb with a SiP modulator
- Bus of 20 cascaded modulators demonstrate scalability
PINE Phase 2 T2M

- Our **industry partners** have shown their commitment to the PINE technology and vision by continuing and increasing their involvement in the PINE Phase 2.

- **Transition Path**
  - We are working with PINE Phase 2 partners NVIDIA and LBNL to identify relevant data center and HPC applications that would benefit from the disaggregated architecture
  - **NVIDIA**
    - As a result of collaboration on PINE Phase 1, NVIDIA, in addition to their continuing collaboration on PINE Phase 2 activities, is increasing their internal silicon photonics activities.
    - This path is to build in-house capabilities for fabless design. NVIDIA sees direct use for PINE technologies in the GPU portfolio, in particular with respect to MCM integration for high bandwidth interconnects.

- **Quintessent** - start-up based on ARPA-E technology, is now a PINE Phase 2 partner continuing their contributions based on hybrid quantum dot technologies

- **Start-ups**, based on PINE technology
  - Columbia comb laser - Xcape to appear in Fast Pitch Session
  - Quintessent now PINE Phase 2 partner, UCSB quantum dot based lasers and SOAs
PINE publications (Q10 forward)

Journals


Conference proceedings


Book Chapter