Towards Exascale HPC Systems: Exploiting Advances in High Bandwidth Memory (HBM2) Through Scalable All-to-All Optical Interconnect Architectures

Pouya Fotouhi, Roberto Proietti, Paolo Grani, Mohammad Amin Nazirzadeh, S. J. Ben Yoo

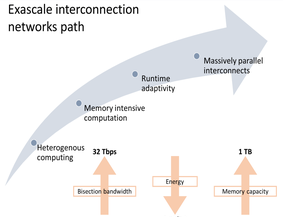
Dept. of Electrical and Computer Engineering, University of California, Davis, CA, 95616, USA

Exascale Challenges

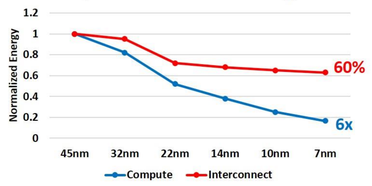
- Data communication between processors and off-chip
  - Consumes lot of energy
  - Experiences high latency[1]
- Flattening the current memory hierarchy through use of:
  - Massive parallel optical interconnects
  - Emerging 2.5D/3D optical interposers
  - High Bandwidth Memory (e.g. HBM and HBM2)
  - Paving the way for memory-driven computing[2]

[1] [2]

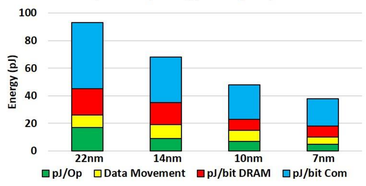
Exascale interconnection networks path



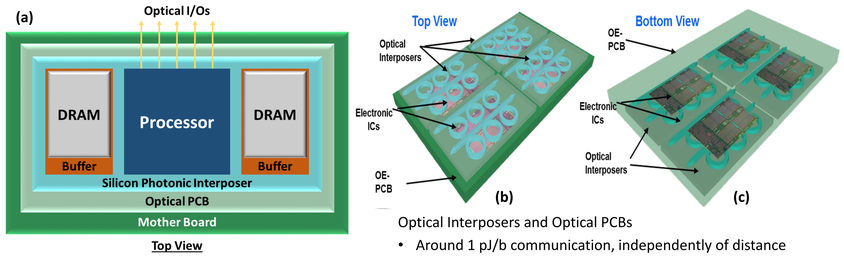
Compute VS On-die Interconnect Energy



Compute Loop Energy Scaling Projections

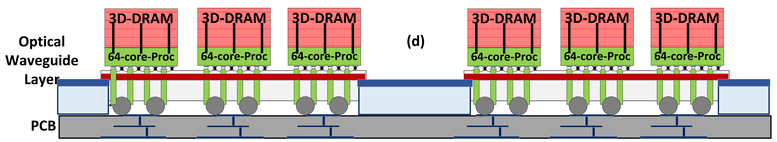


2.5D/3D Integration with Optical Interposer



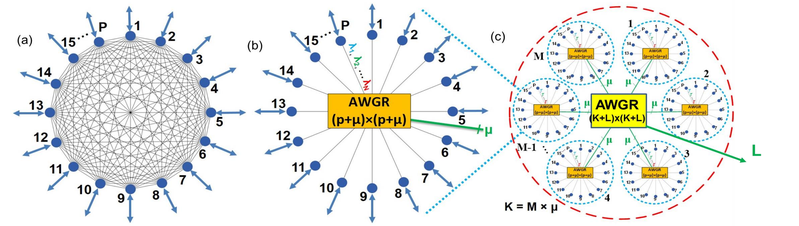
Optical Interposers and Optical PCBs

- Around 1 pJ/b communication, independently of distance
- Flattening the hierarchy and topology



(a) Top view of an OE-PCB containing multiple silicon photonic optical interposer and electronic ICs, and (b) exposed layers top view (from optical interposer side) and (c) its back view (from the OE-PCB side). (d) Multiple many-core processors with 3D-DRAM optically interconnected via optical interposer and optical PCB.

Enabling Massive Parallel All-to-All Optically Interconnected Systems by Arrayed Waveguide Grating Router



One-to-one correspondence between Input and Output Waveguide; single AWGR scalability to 1024 x 1024

Arrayed Waveguide Grating Router (AWGR)
can provide All-to-All communication among N nodes by wavelength routing with N wavelengths and 2N fibers

Silicon Photonic AWGR

Single AWGR Scalability to 512 x 512

Fabricated chip consists of two 8 x 8 AWGR optical switches integrated with silicon photonic transmitter and receiver arrays.

Benchmarking an Optically-interconnected HBM System

As presented in HPCA 17 [1], we studied

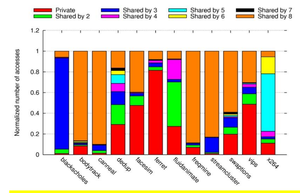
- an optically interconnected system
- With 16 nodes
  - 8CPUs (8 cores each)
  - 8 HBMs attached to the optical switch
  - Each node comprises an HUB for E/O, O/E
- Using gem5 [2] simulator
  - Full-System mode
  - 64 cores running
  - Architectural parameters:

[1] [2]

Cores	64, 64 bits, out-of-order, 2.5 GHz
L1 caches	16kB (I)+16kB (D), 2-way (D/I), 1 cycle
L2 cache	256 kB banks, 8-way, shared, 3/12 cycles tag/tag+data
Directory	MOESI coherence protocol, 64 slices, 1 cycle
E-Interposer network	Concentrated Mesh (8:1 concentration), 2.5 GHz, 64 bits, 1 cycle/hop
O-Interposer network	All-to-All (AWGR-based), 32/5 GHz, 8/17 bits, 1 cycle/hop
Memory	64 GB, 16/8 channels, 128 bits, 1.6 GHz, ~200 cycles request latency

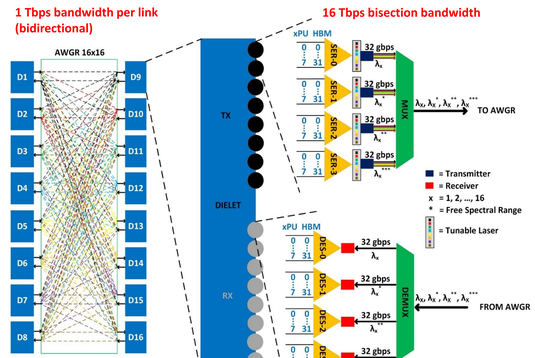
- Running PARSEC [3] benchmark
  - Entire ROI execution time
  - Heterogeneous applications (the graph indicates the sharing traffic pattern)

[3]



1 Tbps bandwidth per link (bidirectional)

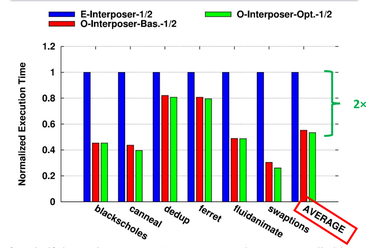
16 Tbps bisection bandwidth



High level view of the Interposer-based network connecting **16 chips** inside the node

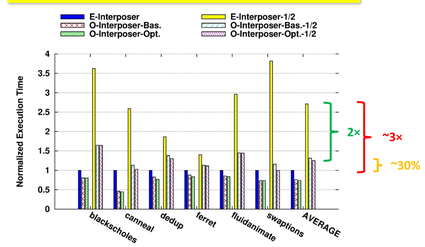
- **How to reach 1Tb/s required for HBM?**
  - We assumed
    - Each processor has 32 8:1 SERDES running at 32Gb/s (4GHz processor)
  - The 32 SERDES were grouped in eight sets of four SERDES
  - Each set of four SERDES
    - Connected optically with HBM modules
    - Through one AWGR using four free-spectral-ranges (FSRs)
  - Therefore, eight AWGRs required to interconnect the processor to the HBM modules

Cache Analysis Results



If we half the cache size, our O-Interposer solutions are still able to achieve good performance exploiting the very low latency (low hops) AWGR-based interconnections. The E-interposer is not able to sustain the increased memory traffic.

Cache Comparison Results



If we half the cache size, the E-interposer performance drops by a factor of ~3× in comparison with the E-Interposer with full cache. Our solution loses only ~30% in comparison with the E-Interposer with full cache, obtaining a ~2× improvement against E-interposer-1/2.

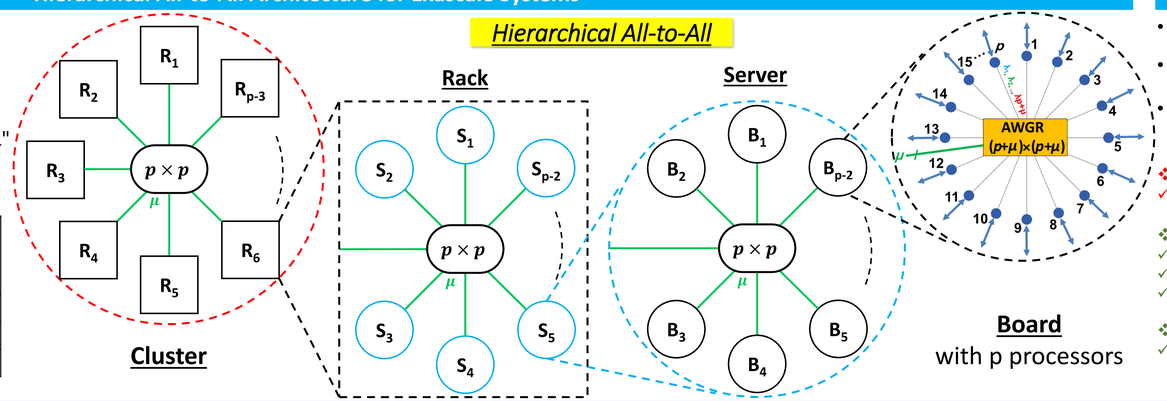
Hierarchical All-to-All Architecture for Exascale Systems

Scalability

- *p:* number of processors in one board
- μ: number of AWGRs for interboard communication
- the total number of SERDES and lambdas per processor is "p+μ"
- μ value affects the total bandwidth to/from the board

p	N_{p_S} = # of processors in one Server	N_{p_R} = # of processors in one Rack	N_{p_C} = # of processors in one Cluster	N_{p_T} = # of TOTAL processors
16	240	3,360	43,680	524,160
18	306	4,896	73,440	1,028,160
20	390	6,840	116,280	1,860,480

Hierarchical All-to-All



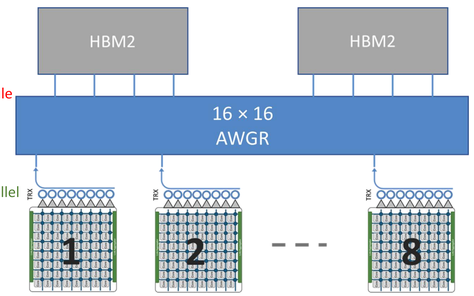
Current Architecture Studies with HBM2

- The architecture discussed above required 32 high-speed TRXs per processor to achieve 1 Tb/s
- With HBM2 bandwidth requirements (2 Tb/s,) the number of TRXs should be doubled
- This study aims investigating trade-offs between:

❖ **Link bandwidth**
✓ Processors sharing the total bandwidth of HBM module

❖ **Complexity**
✓ Avoiding use of Contention Resolution Unit
✓ Fixed-wavelength lasers instead of tunable lasers
✓ A single AWGR as opposed to multiple AWGRs in parallel

❖ **Power consumption**
✓ Far fewer SERDES units required



This work was supported in part by DoD award # W911NF-13-1-0090, DoD contract # H98230-16-C-0820, and NSF grant # 1611560.

NEXT GENERATION NETWORKING SYSTEMS LABORATORY

Learn more about us!