CSR: Small: Collaborative Research: Leveraging Intra-chip/Inter-chip Silicon-Photonic Networks for Designing Next-Generation Accelerators

Amir Kavyan Ziabari, Saiful Mojumder, Yifan Sun, José L. Abellán, John Kim, Ajay Joshi (PI), and David Kaeli (PI)

BOSTON UNIVERSITY — Northeastern

Motivation

Modern workloads demand high compute power from GPUs

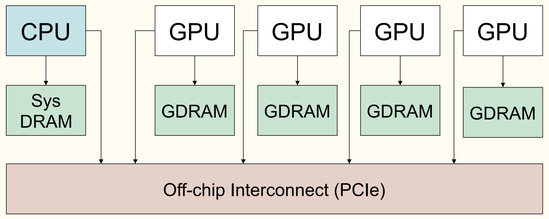
- Big-data applications have high thread-level parallelism and data-level parallelism
- Single main-stream GPUs may not provide sufficient throughput
- Multi-GPU systems are already on the market

AMD Radeon Pro Duo

NVIDIA Tesla V100

Challenges

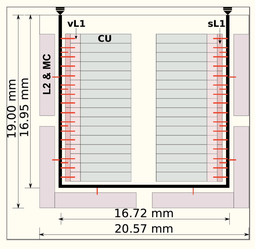
- Memory Management
- Data sharing over low bandwidth bus
  - Compute Units to GPU memory
  - GPU peer to peer
  - CPU to GPU
- High synchronization cost
- Data coherency and consistency



Hybrid Silicon-Photonic NoC [1,2]

Silicon Photonic crossbar using a U-shaped layout

- U-shape lowers the energy and on-chip overhead
- Performance: 2.6x better than electrical crossbar
- Power: 32% higher than electrical crossbar

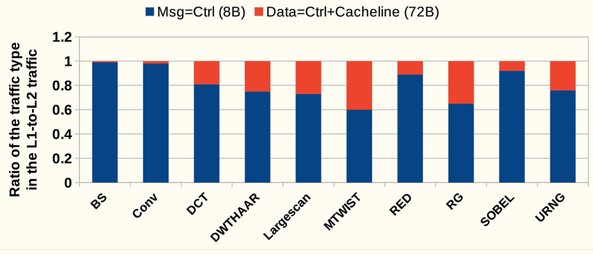


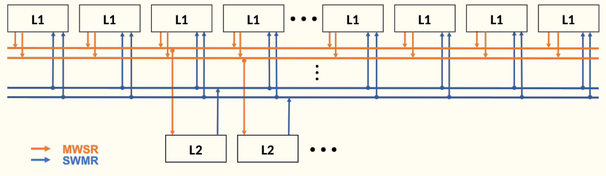
	32-CU		128-CU	
	↑	↓	↑	↓
MWSR	6	40	24	160
SWMR	40	6	160	20
HYB	6	6	6	6

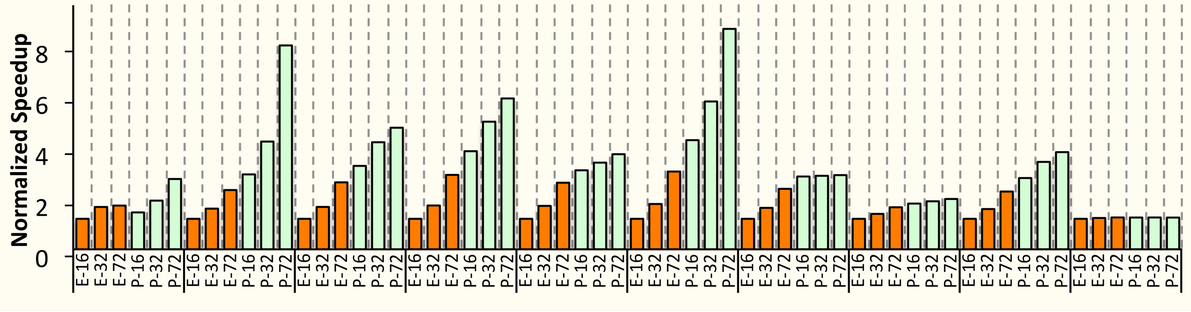
- **Hybrid Design**
  - Multiple-Writer Single-Reader (MWSR) buses for L1-to-L2
  - Single-Writer Multiple-Reader (SWMR) buses for L2-to-L1
  - Scalable design
- **Performance**
  - 2.7X speedup over electrical crossbar
- **Power**
  - Just 2% more than electrical crossbar

Traffic Characteristics

- L2-to-L1 direction
  - Heavy traffic (due to loads), latency sensitive
- L1-to-L2 direction
  - Writes are latency insensitive
  - Dominated by small (8B) control messages
  - Less arbitration latency (no serialization latency)







UMH: A Hardware-Based Unified Memory Hierarchy [3]

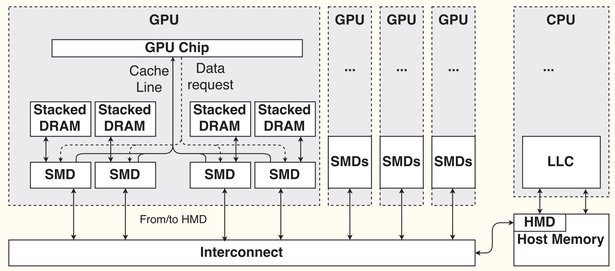
Unified Memory

- Supported since CUDA 6.0, OpenCL 2.0 and HSA
- Improves programmability
- Severe runtime inefficiencies

Unified Memory Hierarchy (UMH)

Extends the memory hierarchy so that the GPU memory acts as a last-level cache for system memory

- No redundant copies
- Remote access only on miss



SMD + HMD

- Allows for GPU data sharing
- Enables CPU coherency

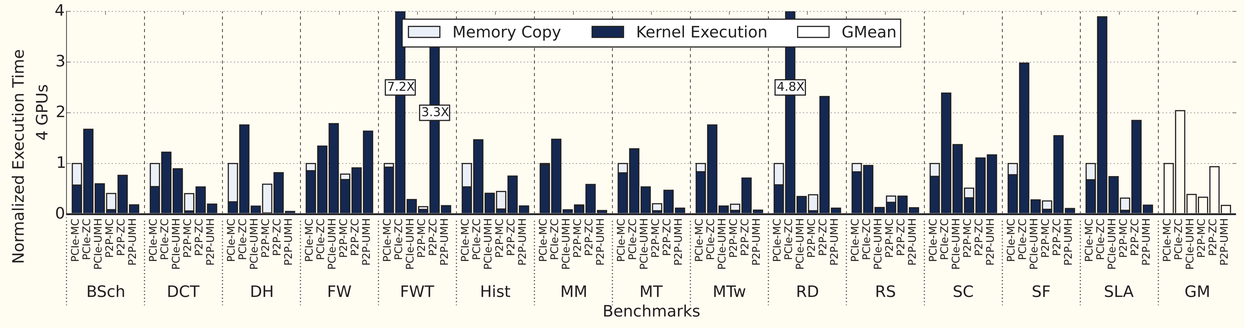
NMOESI cache coherency protocol is used

- Adds the N-state for non-coherent non-exclusive accesses
- 72MB overhead for 1GB data (to support coherence)

Block state	Processor action: *load*	*store*	*n-store*	Incoming request: Eviction	Read request	Write request
N	hit	write request → **M**	hit	writeback → **I**	-	send data → **I**
M	hit	hit	hit	writeback → **I**	send data → **O**	send data → **I**
O	hit	write request → **M**	hit	writeback → **I**	send data	send data → **I**
E	hit	hit → **M**	hit → **N**	→ **I**	send data → **S**	send data → **I**
S	hit	write request → **M**	hit → **N**	→ **I**	-	→ **I**
I	read request → **S** or → **E**	write request → **M**	read request → **N**	-	-	-

Performance

- UMH improves performance by 1.9x on a 4-GPU system, as compared to memcpy
- Going from 1-GPU to 4-GPUs, UMH achieves a 2.3x speedup, while memcpy only achieves a 1.7x speedup



Summary

- A Hybrid Silicon Photonic crossbar provides significantly improved performance, at the cost of slightly higher power consumption.
- UMH improves the performance in a multi-GPU system by eliminating redundant memcpy operations.
- A combination of hybrid Silicon Photonic NoC and UMH can further improve multi-GPU performance.

Future Work

- The performance of a multi-GPU system can be improved further by providing a truly shared memory system.
- The execution time of heterogeneous workloads [4], including machine learning algorithms, can be significantly reduced by introducing shared memory and a high bandwidth interconnection network.

[1] Ziabari, A.-K., et al. *"Leveraging Silicon-Photonic NOC for Designing Scalable GPUs."* Proc. of the 29th ACM International Conference on Supercomputing. ACM, 2015, pp. 273-282.
[2] Ziabari, A.-K., et al. *"Asymmetric NoC Architectures for GPU Systems."* Proc. of the 9th ACM International Symposium on Networks-on-Chip. (NOCS'15), 2015, pp. 1-8.
[3] Ziabari, A.-K., et al. *"UMH: A Hardware-Based Unified Memory Hierarchy for Systems with Multiple Discrete GPUs."* ACM Transactions on Architecture and Code Optimization (TACO) 13.4 (2016): 35, pp. 1-25.
[4] Sun, Y., et al. *"Hetero-mark, a Benchmark Suite for CPU-GPU Collaborative Computing."* Proc. of the IEEE International Symposium on Workload Characterization." 2016, pp. 1-10.