Integrating Nanophotonics in GPU Microarchitecture

Nilanjan Goswami, Zhongqi Li, Ajit Verma, Ramkumar Shankar and Tao Li

Intelligent Design of Efficient Architectures Laboratory (IDEAL), University of Florida

Introduction

GPUs are becoming massively parallel processors

- 35 supercomputers use GPUs (3 in Top10)

GPU performance and interconnect

- Thousands of threads execute simultaneously
- Myriad NoC fetch requests are generated
- Exponential demand of NoC bandwidth and latency

GPU power and heating issues

- ITRS projects 80% chip power goes to NoC

Major advancements in silicon nanophotonics

- Dense wavelength division multiplexed communication fast links
- Minimal dynamic energy consumption

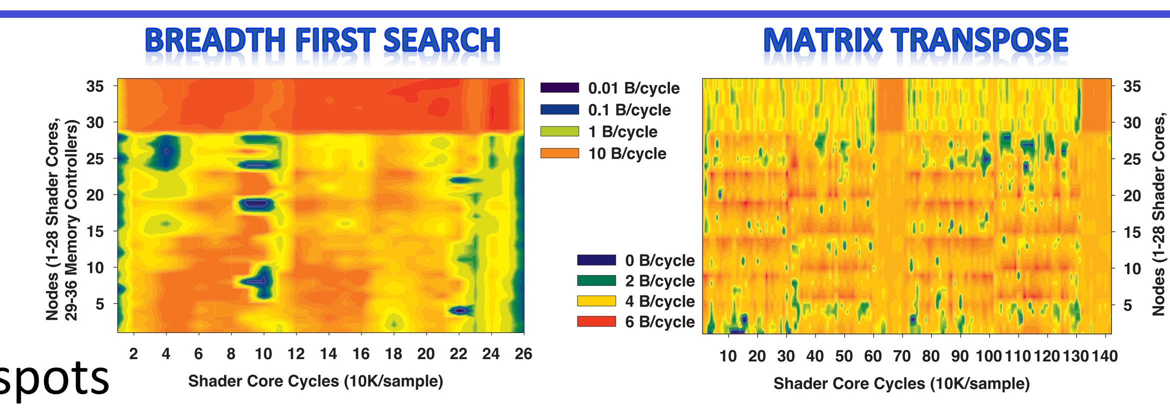
Motivation

Emerging GPGPU workloads

- Increasing computation loads on shaders
- Exert more data traffic on NoC

Bursty network traffic

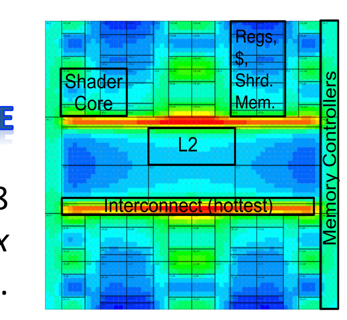
- Data intensive benches causes traffic hotspots



Power and temperature hotspots

- NoC is one of the hottest components

GPU TEMPERATURE PROFILE

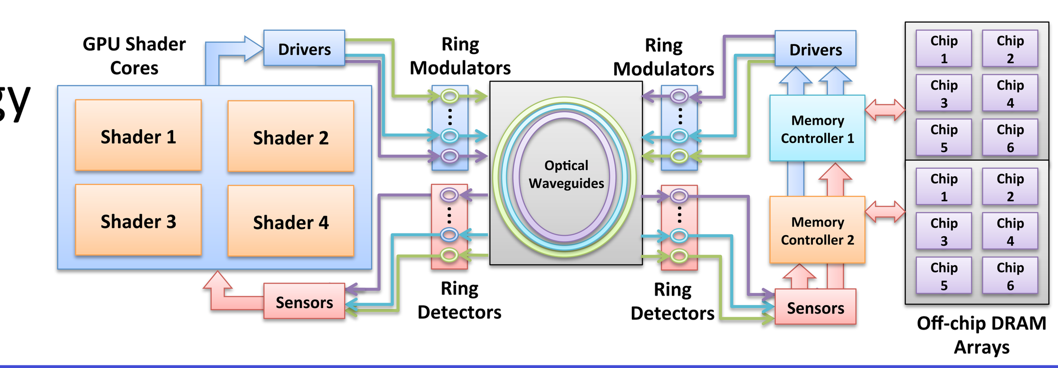


GPU with 16 shader cores and 8 memory controllers running *matrix transpose* workload.

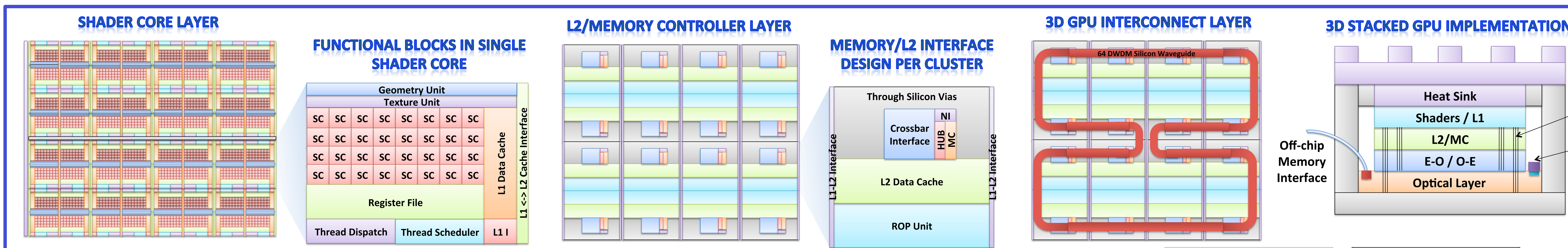
Bursty network traffic exacerbates overall network performances!

Resolution:

- Energy efficient silicon nanophotonic technology
- Fast on-chip communication using DWDM links
- Interconnect aware thread scheduling



Nanophotonic GPU Design

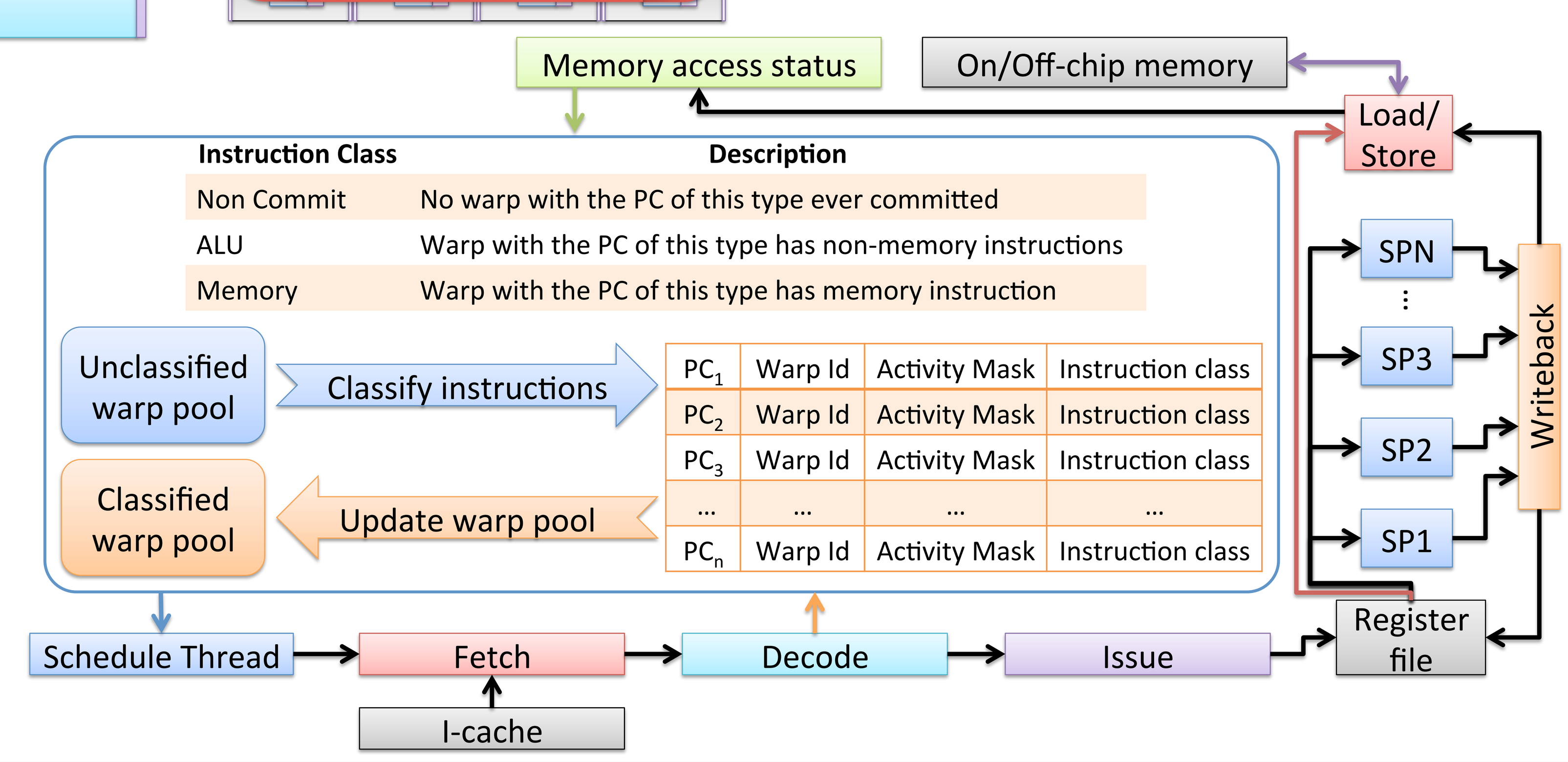


3D stacked GPU with layers for shader core, caches, network, memory interface

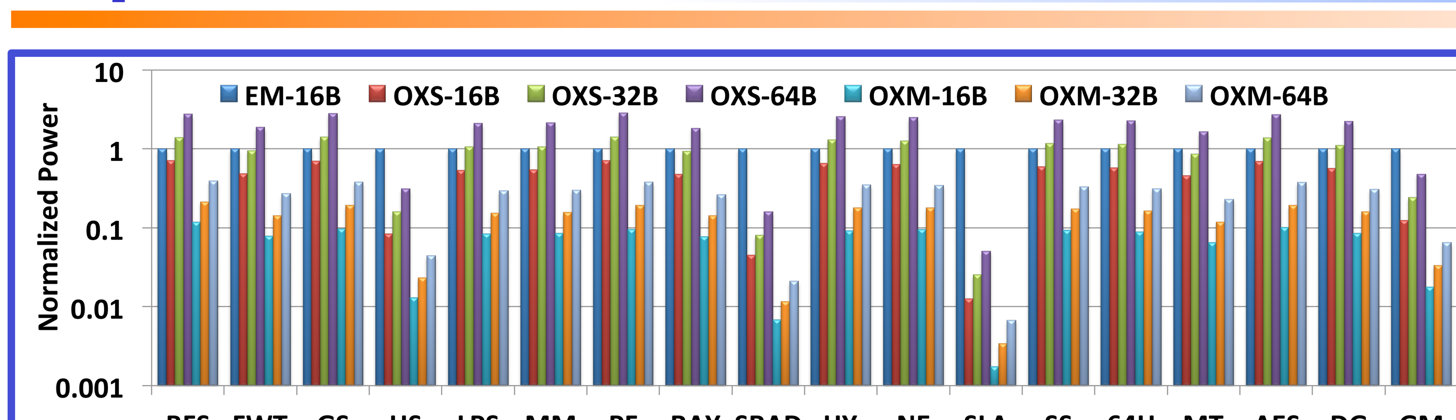
- Network layer has 64 DWDM based SWMR/MWSR crossbar network
- At 5GHz core clock, 16 channels achieve 40Tb/s
- Static power of crossbar is 0.33Watts
- Dynamic energy of crossbar 200fJ/b

Interconnect aware thread scheduling

- Memory accesses are categorized by thread-batch classification of PC
- For known classes of PC, decide which thread-batch to schedule if network is congested

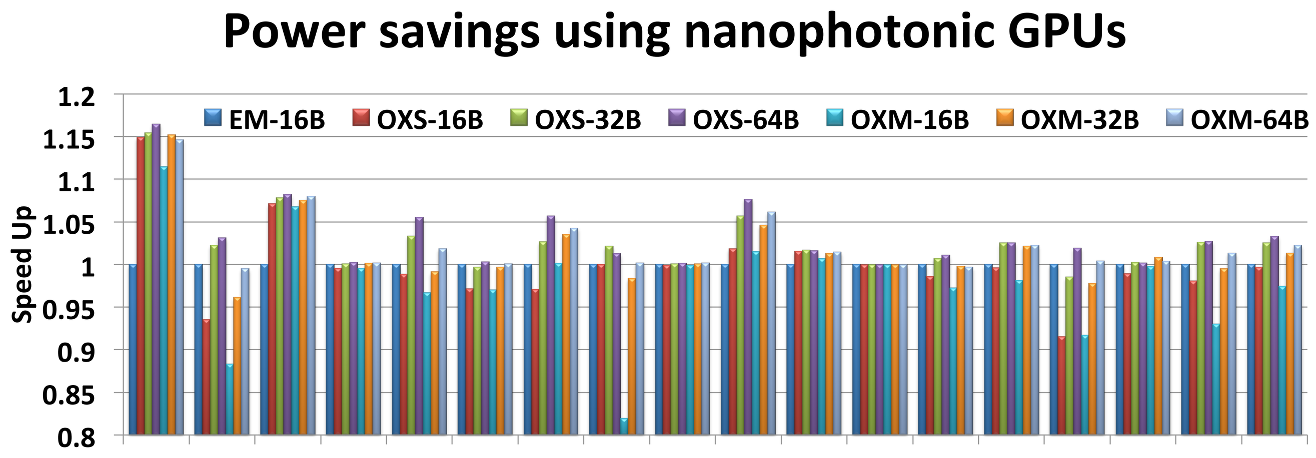


Experimental Results



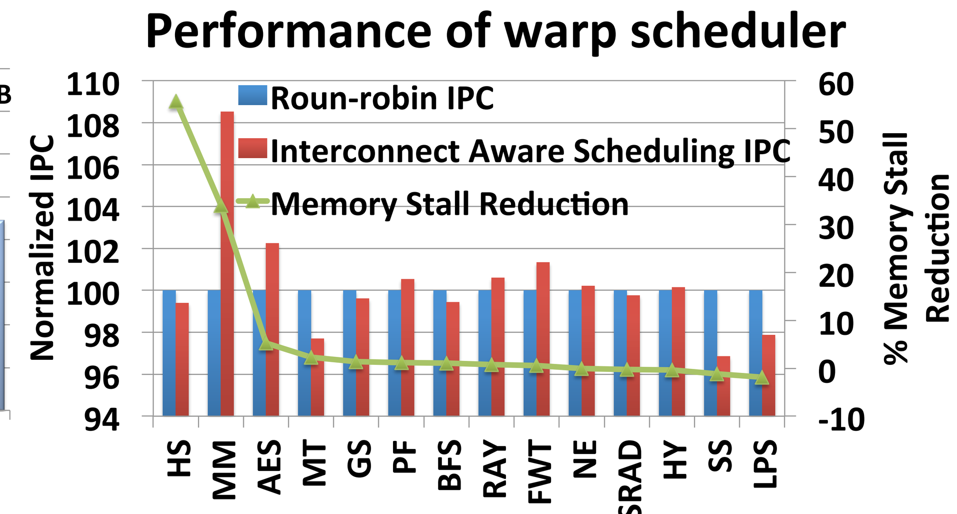
Power savings using nanophotonic GPUs

NoC Types	Description	Channel BW
EM-16B	Electrical mesh network	16 bytes
OXS-16B	Optical crossbar SWMR	16 bytes
OXS-32B	Optical crossbar SWMR	32 bytes
OXS-64B	Optical crossbar SWMR	64 bytes
OXM-16B	Optical crossbar MWSR	16 bytes
OXM-32B	Optical crossbar MWSR	32 bytes
OXM-64B	Optical crossbar MWSR	64 bytes



Performance impact of nanophotonic GPUs

Performance of warp scheduler



Conclusions

Are nanophotonic GPUs energy efficient ?

- SWMR saves 87% power, MWSR saves 98%

Performance impact of nanophotonic NoCs

- Average loss: 1% (SWMR), 2% (MWSR) for 16B BW
- Bursty memory intensive workloads benefit; i.e. breadth-first-search improves by 17%

NoC adaptive thread scheduling performance

- Significantly reduces memory miss stall
- Bursty memory intensive workloads benefit; i.e. breadth-first-search (9%), hybrid sort (5%)

Future Work

Mitigating memory power/performance using nanophotonics memory interface in GPU

- Dynamic and static scheduling exploration

Exploration of overall system behavior

- Power-performance co-optimization

References

A. Bakhoda, G. L. Yuan, W. W. L. Fung, H. Wong, Tor Aamodt, **Analyzing CUDA workloads using detailed GPU simulator**, in Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (**ISPASS**), April 2009

IDEAL RESEARCH
NSF
HPC Center of the University of Florida
UF UNIVERSITY of FLORIDA
The Foundation for The Gator Nation