Characterizing Memory Bottlenecks in GPGPU Workloads

Saumay Dublish
saumay.dublish@ed.ac.uk

Vijay Nagarajan
vijay.nagarajan@ed.ac.uk

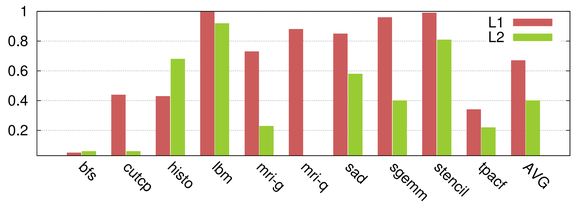
Nigel Topham
nigel.topham@ed.ac.uk

THE UNIVERSITY of EDINBURGH

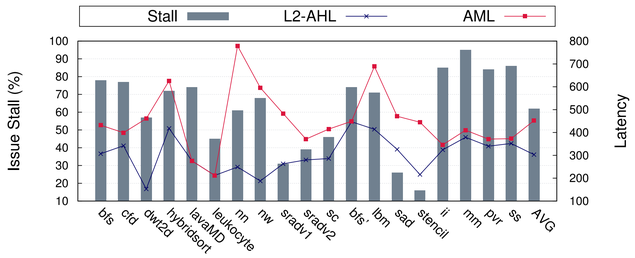
informatics

Introduction

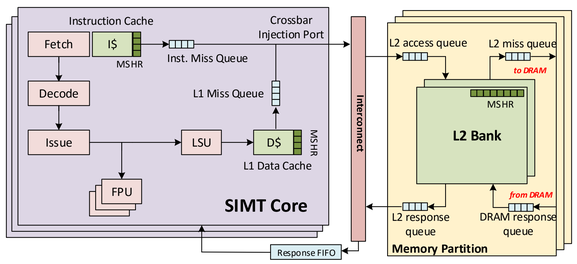
- **Bandwidth demand.** Due to high levels of multithreading, GPGPU workloads present a high demand on the off-chip memory bandwidth.
- **Adding the Cache Hierarchy.** Recent GPU architectures employ a cache hierarchy to filter the off-chip bandwidth demand.
- However, due to high cache miss rates and cache thrashing, the off-chip memory bottleneck is only partly mitigated.



- **Congestion.** The cache hierarchy exposes its own bandwidth limitations in sustaining such high levels of memory traffic, resulting in congestion.



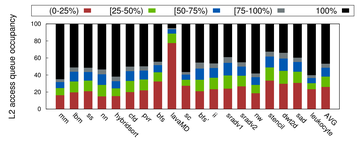
- High memory latencies to the shared L2 cache and off-chip memory indicates severe bandwidth bottleneck and congestion across the entire memory hierarchy.



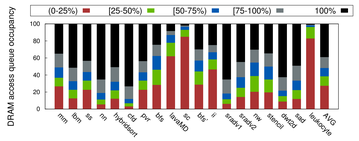
- The scattered nature of the bandwidth bottleneck in GPUs motivates us to analyze the bandwidth implications of the memory hierarchy as a whole.

Motivation and Problem Statement

- Occupancy histograms of buffers between the different levels of the memory hierarchy indicates the level of congestion between adjacent levels.



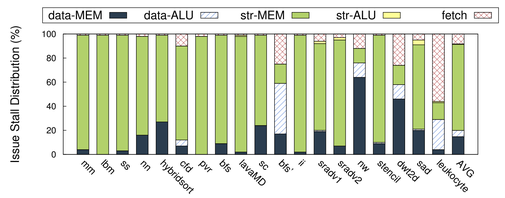
(a) Congestion between L1 and L2.



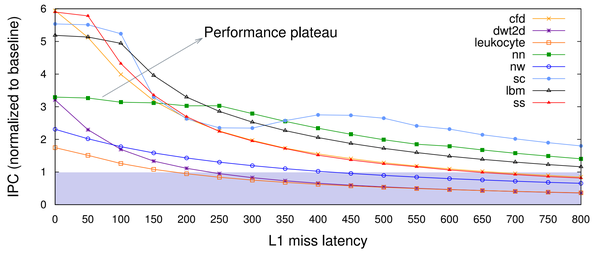
(b) Congestion between L2 and DRAM.

- On average, the access queues to L2 are full for 46% of their usage lifetime and DRAM access queues are full for 39% of their usage lifetime.
- High congestion in the memory system has three major implications.
  - ❶ High latencies appear in the critical path.
  - ❷ Prolonged contention of cache resources.
  - ❸ Back pressure throttling due to a congested lower level.
- Therefore, in this work we aim to characterize and understand the severity of the bandwidth problem posed by the three levels of the memory hierarchy.

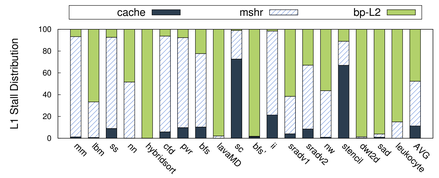
Causes for Core Stalls. The following figure characterizes the core stall cycles due to data, fetch and structural hazards.



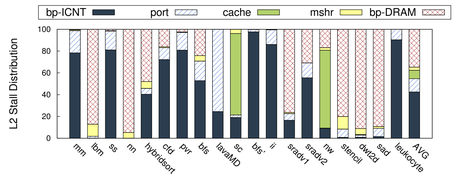
Scope for Improvement? The baseline performance is far from saturation with respect to memory latencies. Therefore, there lies a significant opportunity to improve performance by reducing congestion related latencies.



Stalls in the Memory System

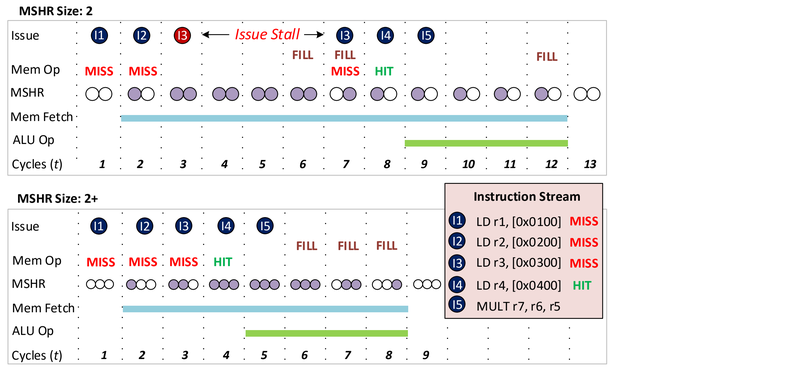


(a) Stalls in the L1 cache.



(b) Stalls in the L2 cache.

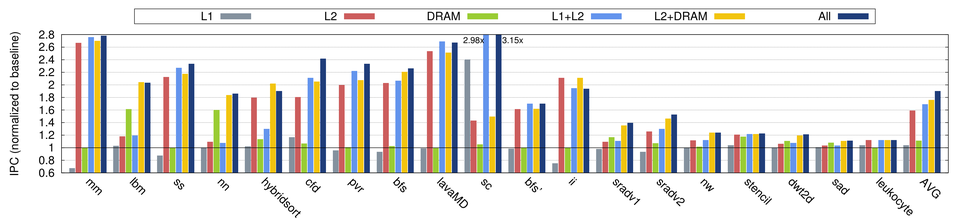
Figure : An example illustrating structural hazard due to lack of MSHR entries.



Key Observations and Conclusion

Memory Level	Architectural Parameters
L1 cache	L1 Miss Queue, MSHR (L1D), Memory Pipeline Width
L2 cache	L2 Miss Queue, Response Queue, MSHR (L2), Access Queue, Data Port, Flit Size, Banks
DRAM	Scheduler Queue, DRAM Banks, Bus Width

- **Criticality of the Cache Hierarchy.** Performance improvement achieved by mitigating the bandwidth bottleneck in the cache hierarchy can exceed the speedup obtained by a memory system with a baseline cache hierarchy and high bandwidth off-chip memory.



- **Synergistic Bandwidth Scaling.** Addressing the bandwidth bottleneck in isolation at specific levels can be sub-optimal and can even be counter-productive. Therefore, it is imperative to resolve the bandwidth bottleneck synergistically across different levels of the memory hierarchy.