Adaptive Cache Management for Energy-efficient GPU Computing

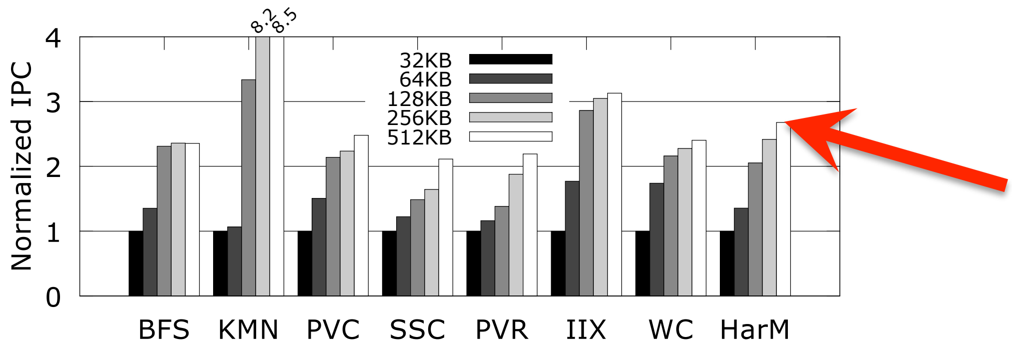
Xuhao Chen[1,2,3], Li-Wen Chang[3], Chris Rodrigues[3], Jie Lv[3], Zhiying Wang[1,2], Wen-Mei Hwu[3]

[1] State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, China

[2] School of Computer, National University of Defense Technology, Changsha, China

[3] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, USA

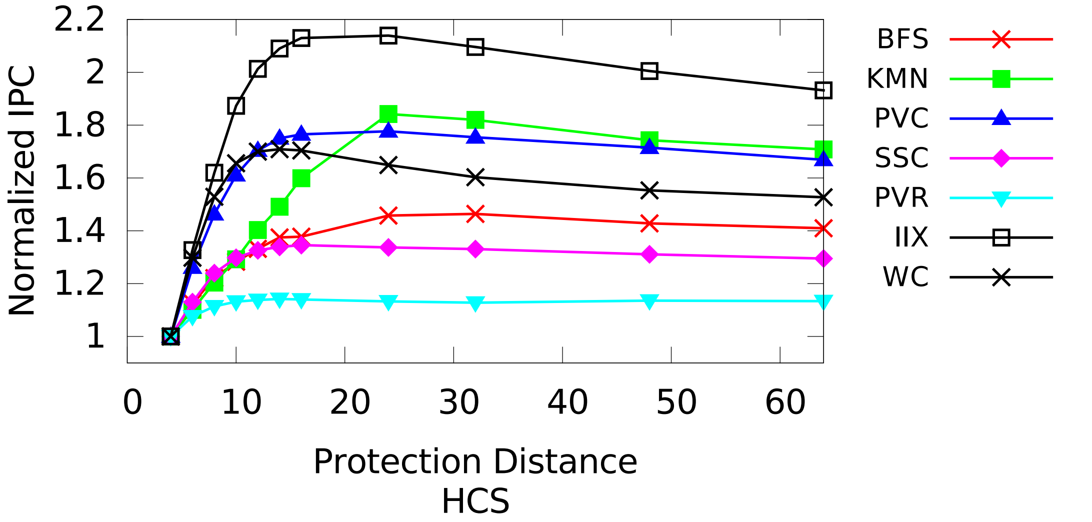
- Many cache sensitive GPU applications have severe cache contention → low cache efficiency → poor performance
  - Smaller L1 cache capacity per thread
- Existing management schemes have limitations
- We propose **Coordinated Bypassing and Warp Throttling (CBWT)** to improve GPU cache efficiency
  - Reduce cache contention rate and NoC latency



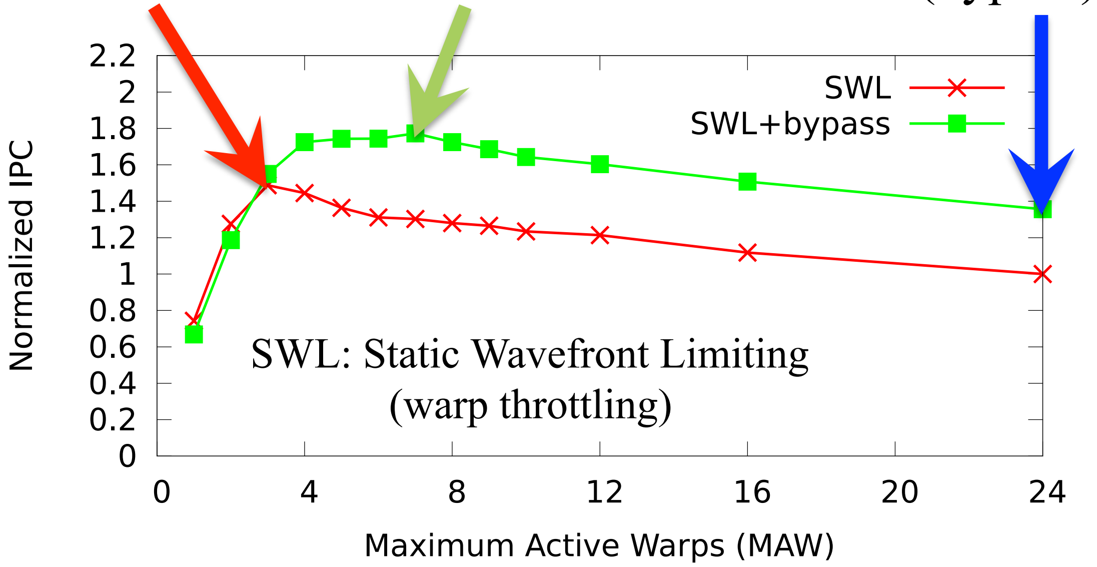
A **2.68x** speedup on average (harmonic mean) for highly cache sensitive (HCS) benchmarks

Observations – Understanding the Limitations

- **Cache bypassing** retains useful cache lines instead of replacing upon miss
  - ✓ Retain useful data → fewer cache misses per thread
    - ✧ Average HCS speedup **1.57x**
  - X High demand on NoC to serve misses, i.e. congestion
  - X Still cannot avoid locality loss
- **Warp throttling** temporarily deactivates some threads
  - ✓ Fewer threads → more cache per thread → fewer misses
  - X Few threads → cannot hide latency through multithreading
  - X Resource under-utilization



1.46x (SWL) **1.84x** (SWL+bypass) **1.38x** (bypass)

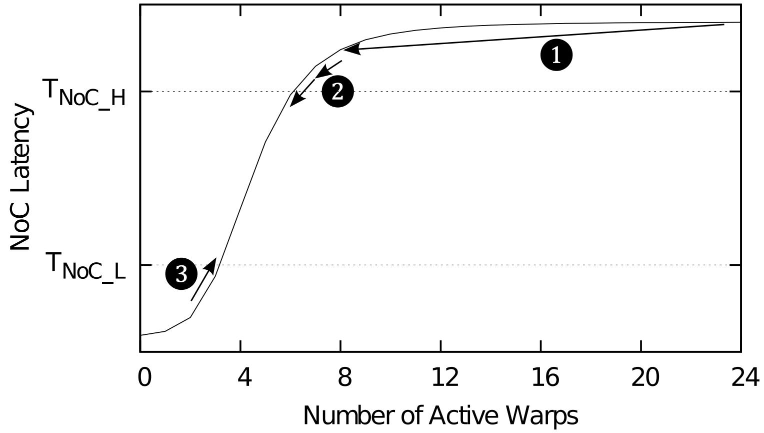


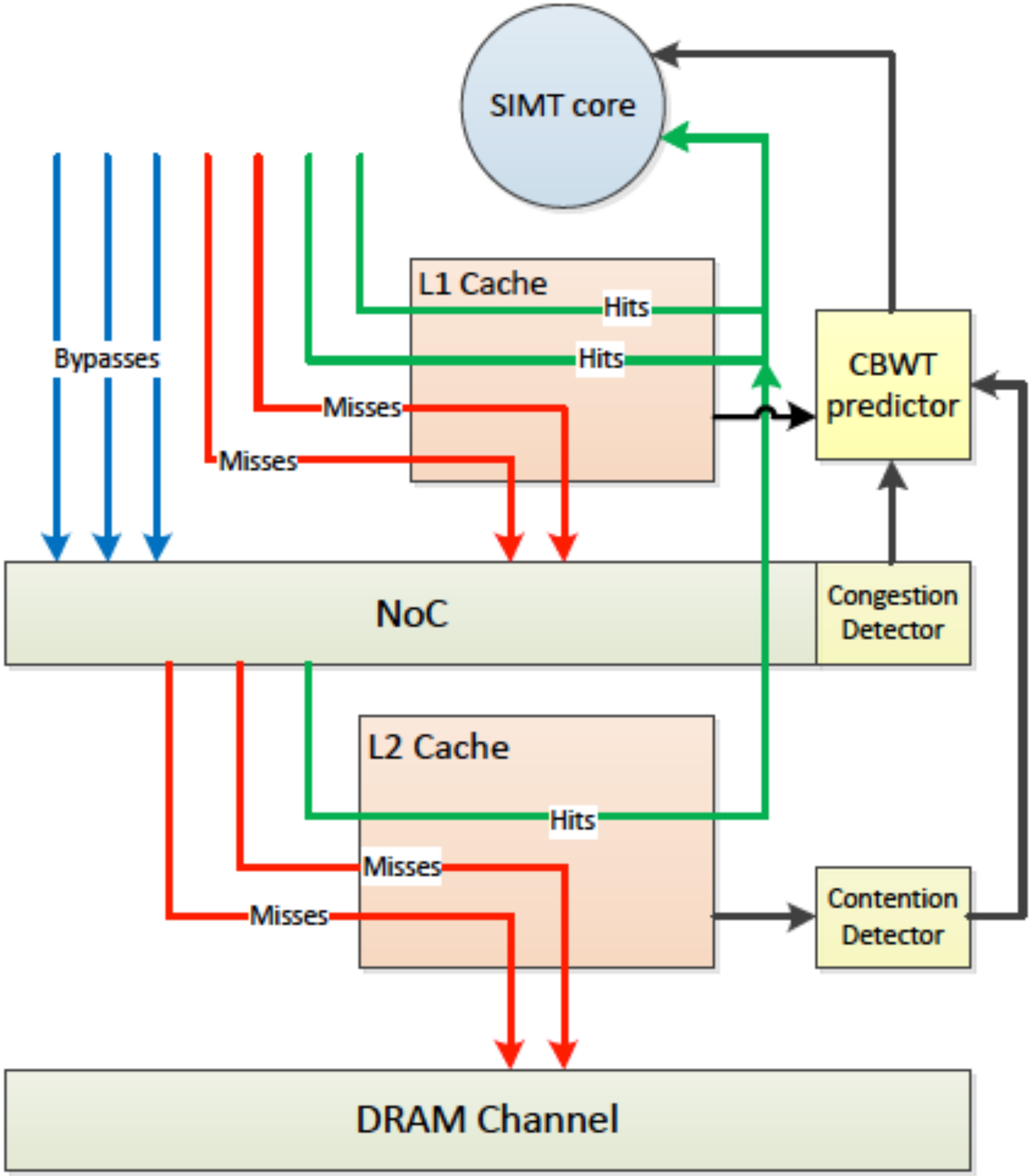
SWL: Static Wavefront Limiting (warp throttling)

CBWT Architecture Overview

- Extra sampling modules (yellow blocks) are added to monitor *contention* and *congestion*.

- Adjust the MAW to keep the network in a *busy* but *low-congestion* range.





Performance and Energy-efficiency

- CBWT achieves an average of **74%** (maximum **661%**) IPC improvement on HCS benchmarks over baseline, which significantly outperforms PDP bypassing (42%) and Best-SWL (52%).
  - PDP bypassing: pure cache bypassing
  - Best-SWL: pure warp throttling

- CBWT outperforms the baseline with an average of **58.6%** Perf/Watt improvement
  - On average, PDP bypassing can reduce 16.5% of DRAM traffic,
  - CBWT reduces DRAM traffic by 54.9%

- Welcome to Session 4A in Main Auditorium on Dec. 16 (Tuesday) at 11:10 AM for more details