Yang Zhang, Zuo-cheng Xing, Cang Liu, Chuan Tang, 2018. CWLP: coordinated warp scheduling and locality-protected cache allocation on GPUs. *Frontiers of Information Technology & Electronic Engineering*, 19(2):206-220. https://doi.org/10.1631/FITEE.1700059

CWLP: coordinated warp scheduling and locality-protected cache allocation on GPUs

Key words: Locality; Graphics processing unit (GPU); Cache allocation; Warp scheduling

Corresponding author: Yang ZHANG
E-mail: zhangyang@nudt.edu.cn
ORCID: http://orcid.org/0000-0001-5919-918X

Motivation

1. Data locality has become increasingly important in designing high throughput and energy efficient GPUs.

2. Preserving data locality becomes especially important for applications.

3. Memory hiding is another important method for improving the GPUs performance. In GPUs, multi-threading is used to hide the long latency and to achieve a high throughput.

Main idea

1. We analyze the data locality in L1 cache for two typical programs. We show that early eviction impairs the performance, and focus on a way to preserve locality to solve the problem.

2. We propose a novel locality detector called the 'instruction PC (LPC)', which is based on the instruction PC, and a prioritised cache allocation unit which evicts the cache line with fewer reuse possibilities, by using the collected reuse information and time-stamp information.

3. A novel locality-based warp scheduler is proposed to use the reuse information from the locality detector to instruct the warp reordering scheme to preserve locality and hide latency at the same time.

Method

1. Locality protected method based on program counter information.

2. Warp reordering method based on locality information.

3. Hardware implementation of coordinated warp scheduling and locality-protected cache allocation scheme.

Major results

Miss rate for the L1 cache with the LPC method

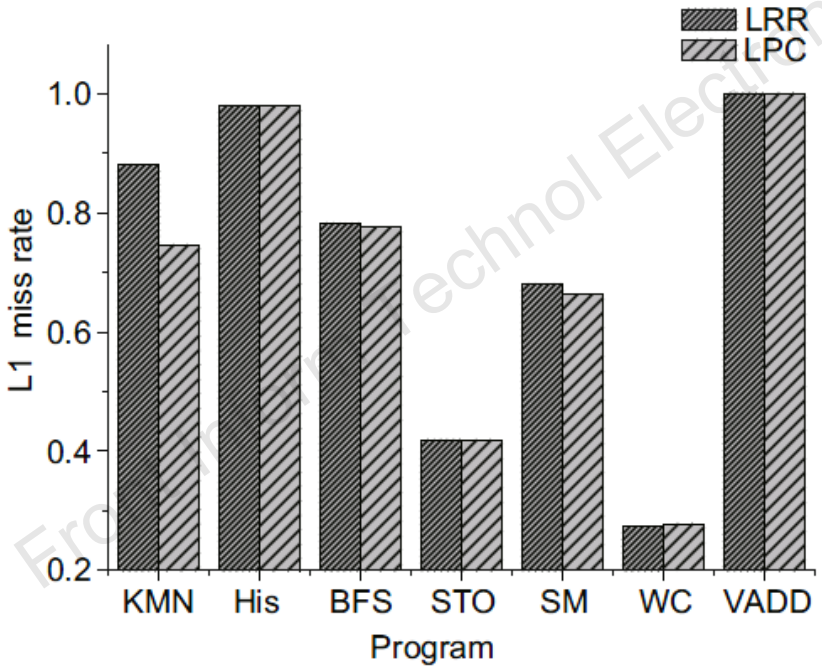

Fig. 10 Miss rate for the L1 cache with locality-protected method based on the instruction program counter (LPC) and loose round robin (LRR) for different programs

Major results

CWLP performance compared with other methods

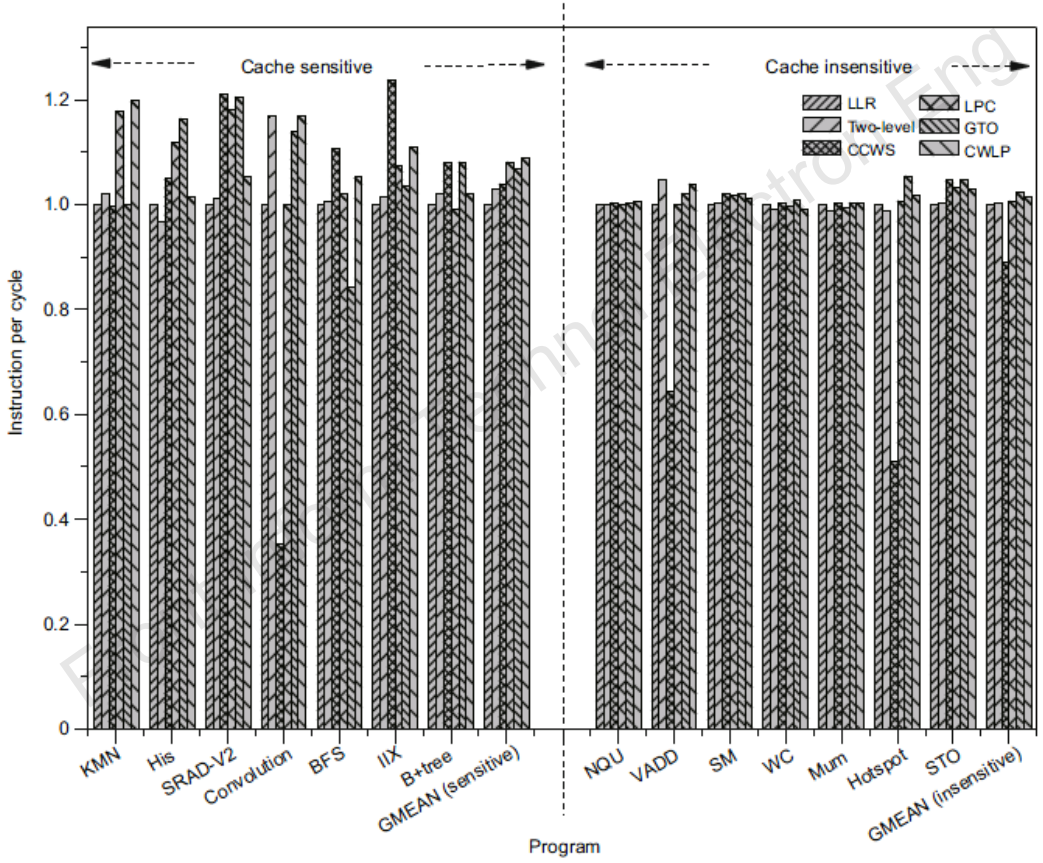

Fig. 11 The CWLP cache allocation scheme performance compared with the LRR, two-level, LPC, CCWS, and GTO schemes for different programs

CWLP: coordinated warp scheduling and locality-protected; LPC: locality-protected method based on the instruction program counter; CCWS: cache conscious wavefront scheduling; GTO: greedy then oldest

Conclusions

1. We Collected the reuse information from each memory block based on the instruction PC and designed a locality predictor to predict the possibility of eviction for each line.

2. We proposed a coordinated cache line evictor which coordinates reuse information with an LRU replacement scheme to evict cache blocks without reuse possibility.

3. We used the locality information to instruct the warp scheduling process to hide latency and preserve locality.