Dynamic Warp Resizing: Analysis and Benefits in High-Performance SIMT
Ahmad Lashgar, Amirali Baniasadi, Ahmad Khonsari
1School of ECE, University of Tehran 2ECE Department, University of Victoria 3School of Computer Science, Institute for Research in Fundamental Sciences

Background
- GPUs are NoC of SIMD cores (SM) and Memory Controllers
- Each SM is a deep-multithreaded processor (1k thread/SM)

Warp Size: Small vs. Large
- Threads are grouped into coarser schedulable elements (warps)
- Benefits: Memory access Coalescing Improving SIMD efficiency
- Disadvantages: Branch divergence Memory divergence

Motivation
- Large warps improve memory access coalescing
- Small warps reduce synchronization overhead, memory divergence, and branch divergence under different warp sizes. IPC is normalized to a GPU using 16 threads per warp.

ISA Modification
- Baseline instruction sequence
- DWR instruction sequence

Experimental Results
- Each configuration of DWR is denoted by DWRs where x denotes the largest warp size within the configuration (e.g., DWR4: 4-way SIMD)
- Benchmarks Characteristics: Large-warps intensive instructions (LAT) show the number of LATs and the number of ignored LATs under DWR (with maximum warp size of 4).

Abstract
- Modern GPUs synchronize threads grouped in a warp at every instruction. These results in improving SIMD efficiency and makes sharing fetch and decode resources possible. The number of threads included in each warp (or warp size) affects divergence, synchronization overhead and the efficiency of memory access coalescing. Small warps reduce performance penalties associated with branch and memory divergence at the expense of a reduction in memory coalescing. Large warps enhance memory coalescing significantly but also increase branch and memory divergence. Dynamic workload behavior, including branch/memory divergence and coalescing, is an important factor in determining the warp size returning best performance. We propose Dynamic Warp Resizing (DWR) to adjust warp size during runtime and according to program characteristics.

Contact
Ahmad Lashgar
a.lashgar@ece.ut.ac.ir
Amirali Baniasadi
amirali@ece.uvic.ca
Ahmad Khonsari
ak@ipm.ir