Unified On-chip Memory Allocation for SIMT Architecture

Ari Hayes
Computer Science

12/21/2018 at 02:00 pm
Hill 350

Abstract

The popularity of general purpose Graphic Processing Unit (GPU) is largely attributed to the tremendous concurrency enabled by its underlying architecture – single instruction multiple thread (SIMT) architecture. It keeps the context of a significant number of threads in registers to enable fast "context switches" when the processor is stalled due to execution dependence, memory requests and etc. The SIMT architecture has a large register file evenly partitioned among all concurrent threads. Per-thread register usage determines the number of concurrent threads, which strongly affects the whole program performance. Existing register allocation techniques, extensively studied in the past several decades, are oblivious to the register contention due to the concurrent execution of many threads. They are prone to making optimization decisions that benefit single thread but degrade the whole application performance.

Is it possible for compilers to make register allocation decisions that can maximize the whole GPU application performance? We tackle this important question from two different aspects in this paper. We first propose an unified on-chip memory allocation framework that uses scratch-pad memory to help: (1) alleviate single-thread register pressure; (2) increase whole application throughput. Secondly, we propose a characterization model for the SIMT execution model in order to achieve a desired on-chip memory partition given the register pressure of a program. Overall, we discovered that it is possible to automatically determine an on-chip memory resource allocation that maximizes concurrency while ensuring good single-thread performance at compile-time. We evaluated our techniques on a representative set of GPU benchmarks with non-trivial register pressure. We are able to achieve up to 1.70 times speedup over the baseline of the traditional register allocation scheme that maximizes single thread performance.

Examination Committee: Prof. Zheng Zhang (Chair), Prof. Ulrich Kremer, Prof. Manish Parashar, Prof. Konstantinos Michmizos