Compiler-Assisted Workload Consolidation to Efficiently Exploit Dynamic Parallelism for Recursive Applications

Hancheng Wu
Dr. Michela Becchi, Thesis Supervisor

Abstract

GPUs have been widely used to accelerate applications for its high throughput. Since Kepler architecture, Nvidia introduced a new feature-Dynamic Parallelism (DP), which enables the spawning of GPU functions from inside a GPU function. This makes implementing recursion on GPU without the intervention of CPUs possible. However, preliminary data shows that simple DP-based implementations of recursion result in poor performance. This work focus on how to efficiently exploit DP for parallel recursive applications on GPU by performing workload consolidation using compiler techniques.

The evaluation shows that GPU kernels consolidated with the proposed code transformations achieve an average speedup in the order of 1500x over basic implementations using DP and an average speedup of 3.9x over optimized flat GPU kernels for both tree traversal and graph based applications.