Efficient Batched Predecessor Search in Shared Memory on GPUs

Henri Casanova*[1]

[1]Information and Computer Sciences [Hawaii] (ICS) – Pacific Ocean Science and Technology (POST) Building, Room 317 1680 East-West Road Honolulu, HI 96822, United States

Abstract

Many-core Graphics Processing Units (GPUs) are being used for general-purpose computing. However, due to architectural features, for many problems it is challenging to design parallel algorithms that exploit the full compute power of GPUs. Among these features is the memory design. Although the issue of coalesced global memory access has been documented and studied extensively, another important architectural feature is the organization of shared memory into banks. The study of how bank conflicts impact algorithm performance has only recently begun to receive attention. In this work we study the predecessor search algorithm and the effects of bank conflicts on its execution time. Via complexity analysis we show that bank conflicts cause significant loss in parallelism for a naive algorithm. We then propose two improved algorithms: one that eliminates bank conflicts altogether but that uses a work-inefficient linear search, and one that is work-optimal but that experiences a limited number of bank conflicts. We develop GPU implementations of these algorithms and present experimental results obtained on real- world hardware. These results validate our theoretical analysis of the naive algorithm and allow us to assess the performance of our algorithms in practice. Although both our improved algorithms outperform the naive algorithm, our main experi- mental finding is that our conflict-limited algorithm provides a larger performance gain.

*Speaker