Qilin: Exploiting Parallelism on Heterogeneous Multiprocessors with Adaptive Mapping

1 Summary

In this paper, the authors proposed adaptive mapping, a fully automatic technique to map computations to processing elements on heterogeneous multiprocessors. They implemented the idea on an experimental heterogeneous programming system called Qilin for programming CPUs and GPUs.

Qilin system itself is a layer between application source code and hardware, it consists of a dynamic compiler and its code cache, a number of libraries, a set of development tools, and a scheduler. It is written in c/c++ and provide an API to programmers for writing parallelizable operations. By taking advantage of automatically adaptive mapping, programmers are released from a lot of manual mapping work, they are only asked to call Qilin APIs and focus on the logic of program.

The experimental results demonstrate that, for a set of important computation kernels, automatic adaptive mapping achieves a speedup of 9.3x, which is 69% and 33% faster than using the CPU or GPU alone, respectively. In addition, adaptive mapping is within 94% of the speedup of the best manual mapping found via exhaustive searching.

2 Intellectual Merit

1. This paper presented adaptive mapping, first introduced the technology of automatically maps computations to processing elements on heterogeneous multiprocessors.

2. This paper use experimental statistic to show illustrate that the optimal mapping from computations to processing elements highly depends on the input problem size and the hardware capability. The distribution of workloads to CPU and GPU would be quite different even for single application. This paper conclude that any static mapping techniques would not be satisfactory, an dynamic mapping technique that can automatically adapt to the run-time environment is necessary.

3. The authors of this paper have implemented it in an experimental system called Qilin for programming CPUs and GPUs. They use experimental evaluations to demonstrate that automated adaptive mapping performs close to manual mapping and can adapt to changes in input problem sizes and hardware configurations.

3 Weakness

1. Qilin is designed for only single application. It did not consider multiple application schedule and mapping, which is a point that can be improved.

2. Qilin doesn’t support several different types of processing devices, such as CPUs with integrated GPUs and multiple discrete GPUs. Their prototype are build using one CPU and one GPU.

3. Qilin system does not work transparently. It need the programmers to use Qilin APIs for writing parallelizable operations.

4. The applicable data parallel kernels are limited by usage of the APIs, which requires access locations of all threads to be analyzed statically.