論文の内容の要旨

Compiler Optimizations for Energy-Efficiency of Heterogeneous Computing Systems

(ヘテロジニアス計算機システムの電力効率向上のためのコンパイラ最適化手法)

氏名 薦田 登志矢

Energy efficiency has become one of the most important metric in recent computing systems. The maximum power consumption of the large scale computing center is reaching to the physical limitation of the power supply system. Now, we have to scale the system performance without increasing the power consumption. To this end, heterogeneous systems with accelerators are becoming popular due to its high performance and its high energy efficiency. The peak performance and the peak energy efficiency of such systems are much higher than those of conventional homogeneous systems. However, achieving high energy efficiency in practical situations is not easy because the energy efficiency of the heterogeneous systems highly depends on the optimizations used in the applications running on them.

To get great performance improvement from accelerators, a large fraction of the application program must be parallelized and optimized to utilize accelerators. This is because the performance improvement is limited to the fraction of the program which is not accelerated, according to Amdahle’s law. To increase the fraction of the program offloaded to accelerators, it is essential to enable programmers to utilize accelerators easily. However, the programmability of the current accelerator platform is very low due to the lack of sophisticated optimization techniques in the compiler.

Even after optimizing large fraction of the program to use accelerators, to draw the maximum energy efficiency from the systems, it is necessary to reduce the power consumption wasted in the system components which do not contribute to the performance improvement. For example, leakage power

consumed both in CPUs and accelerators occupies the large fraction of the total system power consumption. Also, computing devices which have less tasks to do may run at an unnecessarily high frequency and it results in large dynamic power consumption. In conventional homogeneous systems only with CPUs, these unnecessary power consumption has been successfully reduced by some power optimization techniques, such as power gating or dynamic voltage and frequency scaling (DVFS). However, it is unclear that how we should apply these power optimization techniques to the heterogeneous systems with accelerators.

To this end, in this dissertation, we propose three new compiler and runtime optimization techniques which aim to solve the above problems. The three proposed optimization techniques can be orthogonally applied to applications running on the heterogeneous system and each technique independently contributes to improving the energy efficiency of the system.

First, to ease the accelerator programming process, we propose a directive based accelerator compiler which has the capability of utilizing multiple accelerators automatically. The compiler provides single monolithic memory space on top of the disjoint accelerator memories, and programmers can transparently utilize multiple accelerators. The experimental results show that the performances of the proposed system are close (70% – 80 %) to those of hand-written CUDA programs while it can save a lot of programming efforts in utilizing multiple GPUs.

Second, to reduce the leakage power in CPU functional units, which have low utilization because the parallel tasks have been offloaded to accelerators, we propose a compiler directed power gating control technique. Using a static analysis to predict precise length of idle periods in CPU functional units, the proposed sleep control technique can effectively reduce the leakage power consumed in CPU functional units. The experimental result shows that the proposed sleep control technique achieves up to 19 % larger leakage power reduction than that of conventional sleep control techniques.

Finally, to avoid the inefficient execution due to the load imbalance between CPUs and accelerators, we investigate the runtime software technique with which we can utilize both CPUs and accelerators in parallel to execute the same data parallel task. In addition to balancing the load between CPUs and accelerators, the proposed technique can cooperatively control the frequency of CPUs and accelerators to further optimize the energy efficiency of the system. In the proposed technique, empirical models of performance and power are proposed to guide the settings of device frequencies and the heterogeneous task mapping. The experimental result shows that the empirical models are precise enough to select near optimal parameter sets in most cases, and the proposed technique enables the system to achieve the almost ideal performance under a certain power constraint.

From the experimental results about the proposed optimizations, we conclude that the sophisticated compiler and runtime optimization techniques greatly help us achieve high energy efficiency in the heterogeneous systems equipped with accelerators.