PROJECT ABSTRACT

Energy Efficient Computing on GPU-based Heterogeneous Systems

TECHNICAL DESCRIPTION: The current trend in designing future multiprocessors is to integrate hundreds of cores and hardware accelerators (HAs), such as GPUs, on a single platform. However, as the system size scales, the power and energy consumption of these heterogeneous multiprocessors vastly exceed the budget. The primary focus of the project is to develop energy efficient techniques that can be implemented in the GPU with proper coordination with the CPUs. Energy reduction in GPU-based heterogeneous multiprocessors can be achieved by extending current CPU techniques to GPU. The project develops new runtime techniques for GPU core scaling and DVFS, and combines them with basic changes in algorithms and data structures to improve energy efficiency.

Energy Saving in GPU Applications: Existing models for performance and energy consumption assume 100% utilization of the processing cores and perfect overlap with memory access during execution. This project develops a more accurate model for energy efficiency taking into account the algorithm, data structure, caching and memory coalescing for different applications.

Runtime Core scaling and DVFS: A runtime system is being developed that monitors the GPU core and memory utilizations together with the energy consumption while executing an application. The runtime adjusts the number of cores and/or frequency level dynamically through prediction, but continues to make corrections as the execution proceeds. The runtime is extended to heterogeneous systems consisting of both CPU and GPU.

Algorithm Based DVFS: The current DVFS techniques for scientific computing applications cannot fully eliminate slacks, therefore, are not energy optimal. By leveraging the algorithmic characteristics, a frequency scheduling technique is developed for linear algebra applications to achieve better energy efficiency.

Algorithmic Based Task Design: The project optimizes the energy efficiency while partitioning and designing tasks of an application. Without loss of generality, Cholesky factorization is used as an example and an energy efficient scheduler is developed.

BROADER IMPACT: The project develops software products that can be readily applied to existing large scale heterogeneous computers executing scientific applications that are suitable for defense, energy and critical infrastructure projects. Also applications like weather forecasting and structural dynamics are developed that have a great impact on society. The research content is integrated to graduate courses to provide training to students for designing and programming heterogeneous systems. The project aims to produce very high quality Ph.D. graduates including female students. The University of California, Riverside is known for its large proportion of Hispanic students, and UCR is a minority-serving institution. The project supports recruiting underrepresented minority and female students.