Real-Time Computing Using GPUs

Department of Computer Science University of North Carolina at Chapel Hill August 2014

The Challenge

The computing industry recently experienced a major shift in CPU architectures with the advent of multicore chips. This shift has necessitated the adoption of new programming models, algorithms, and analysis methods to fully exploit the parallelism inherent in multicore chip designs. While advances in these areas are well underway, industry has already begun yet another architectural shift towards heterogeneity in order to achieve greater levels of performance and energy efficiency. Heterogeneity creates new challenges because the availability of different types of processing resources means that "choices" must be made when allocating hardware resources to software components. The need to resolve such choices can add considerable complexity to resource allocation. One of the most successful applications of heterogeneity today is in architectures in which powerful graphics processing units (GPUs) are used alongside general-purpose CPUs. Though originally intended as special-purpose graphics accelerators, GPUs are now being widely used for non-graphics processing in many application domains. The breadth of such domains is now expanding to include many applications in which real-time constraints exist. For example, envisioned automated automotive systems will require real-time sensing and tracking features that GPUs can accelerate. Unfortunately, because real-time applications require predictable execution, utilizing GPUs in supporting them is not straightforward. **The goal of the proposed research is to determine which resource allocation methods best facilitate the support of real-time applications on heterogeneous platforms that may have multiple CPUs and GPUs.**

The Approach

This goal will be met by undertaking a broad study of issues affecting the deployment and analysis of real-time applications implemented on GPU-enabled multicore platforms. This project will contribute new real-time resource allocation methods that can be applied when such platforms are used and associated analysis for checking timing constraints. The latter will require addressing a number of difficult analysis issues that arise when GPUs are present. Additionally, working prototypes of the most viable resource allocation methods will be implemented and a large body of experimental evaluation data concerning them will be produced. These experimental efforts will involve both synthetic workloads for which parameters affecting real-time correctness can be controlled and varied, and actual workloads from case-study systems. These efforts will enable GPUs to be exploited in supporting real-time applications to an extent not possible today.

The Significance

Heterogeneous architectures in general, and CPU+GPU combinations in particular, are seeing increasingly widespread use in industry; thus, the insights, algorithms, and software produced in this project could have significant industrial impact. The proposing group has a longstanding tradition of working with industry partners.

Project Members

James Anderson (PI)
Sanjoy Baruah (Co-PI)

Research Sponsor

National Science Foundation (NSF)

For More Information

Dr. James Anderson
Department of Computer Science
University of North Carolina at Chapel Hill
CB #3175, Sitterson Hall
Chapel Hill, NC 27599-3175
Phone: (919) 962-1757
Fax: (919) 962-1799
E-mail: anderson@cs.unc.edu

http://www.cs.unc.edu/~anderson/projects/rtgpu.html