## EXECUTIVE SUMMARY

**PROJECT NAME:** G9-11: Efficient Task Scheduling for Modern GPU-based Heterogeneous HPC Systems

**INVESTIGATOR(S):** Dr. Tarek El-Ghazawi, GWU site of CHREC

### PROJECT DESCRIPTION

Parallel systems with multi/many-core processors accelerated with GPUs are becoming commonplace, due to rapid advances in the programmability and performance of GPU architectures. The GPU acts as a co-processor device, with the performance benefits derived by offloading suitable tasks from the parallel application onto the GPU. However, the GPU can be under-utilized if several of the tasks do not consume all the available GPU resources, when multiple sequential tasks need to be accelerated. Furthermore, data transfer between the CPU and the GPU constitutes a significant overhead that has to be mitigated to the extent possible. The proposed project, G9-11, will focus on efficiently managing the GPU resources and the associated overheads, by exploiting capabilities of modern GPU devices. Beginning with a simple model that attempts to predict the GPU performance in the presence of multiple tasks, the proposed research aims at developing efficient scheduling techniques for choosing the appropriate sequence and set of tasks to be executed on the GPU. By providing a user-driven tool for achieving the required scheduling, the project will culminate with recommendations for a suitable programming interface that can efficiently make use of the developed scheduling infrastructure.

### EXPERIMENTAL PLAN

The proposed solution will be composed of the following stages of development:

(a) *Task profiling and performance modeling*: Carry out a profiling of tasks from representative applications in order to enable prediction of performance and resource usage patterns. This essentially involves the study of the capabilities of modern GPUs, such as concurrent kernel execution and overlapped data transfer and execution.

(b) *Efficient scheduling algorithm*: Investigate and develop scheduling techniques for efficient execution of tasks on a GPU. The developed algorithms will make use of the aforementioned performance model and exploit the modern GPU capabilities. Scheduling will allow offline selection and sequencing of tasks to achieve maximum overlapping among tasks in terms of concurent execution as well as I/O transfers for input and output data.

(c) *User-driven tool and API specification*: Develop a user-driven tool for carrying out the required scheduling, based on user-provided inputs on the available tasks and their profiling parameters. A suitable API will then be suggested, which may then be used for modifying the application to capture the schedule generated by the tool.

(d) *Evaluation*: Explore and develop a methodology for evaluating the proposed task scheduling. The developed infrastructure will exercise the system with some benchmark application tasks to obtain performance measurements and comparisons with other ad-hoc approaches, on latest NVIDIA GPUs.

### HOW THIS PROJECT IS DIFFERENT

This project will provide one of the earliest solutions to support offline scheduling of GPU tasks that are to be executed within an application. Since this support is currently lacking in commercial tools, it will allow developers to make maximize the utilization of the powerful GPU resources available in contemporary HPC systems. The proposed solution will also position us to extend it for heterogeneous compute resources, which has the potential of far reaching impact.

### POTENTIAL MEMBER COMPANY BENEFITS

- Member companies will have access to one of the first scheduling tools for GPU resources in HPC systems.
- In addition to the tool, members will be able to directly use the provided examples and evaluation methodologies for their systems.
- Member organizations will also be able to influence the project direction by providing valuable feedback.

### EXPECTED DELIVERABLES

- A user-driven task scheduling tool for efficient utilization of resources in NVIDIA Fermi GPUs. This will include the reference code as well as usage examples.
- Several conference and journal publications arising out of the proposed research.

### PROJECT BUDGET

3 CHREC Memberships