In this work, we provide a methodology to predict execution time on multiple GPUs, which allows us to determine the optimal number and configuration of GPUs for a given program’s execution. The benefits include:

- Allowing developers to predict the most appropriate GPU configuration for an application without having to purchase hardware or create a full software implementation.
- Alleviating some of the need for hardware-specific tuning, increasing portability and decreasing complexity.
- Providing additional speedup by utilizing more resources.

GPGPU is currently focused on enhancing performance on a single GPU (usually yielding 10X-40X speedup). Although larger speedups are obtainable from a single GPU, they require abandoning general purpose code in favor of graphics-based, architecture-specific algorithms.

Our alternative methodology is a technique for execution prediction and enhancement using multiple GPUs.

Considerations:
- Average 11% difference in prediction versus actual execution time over all applications (max 40%)

Acknowledgement
This work was supported in part by the Institute for Complex Scientific Software, and Gordon-CenSSIS, The Bernard M. Gordon Center for Subsurface Sensing and Imaging Systems, under the Engineering Research Centers Program of the National Science Foundation (Award Number EEC-9986821).

References
