

ENGINEERING SEMINAR SERIES, FALL 2009

Thursday, November 12, NC2607/09, 11:000a.m.

*

High Performance Computing on GPU Processors

Dan Connors

**Department of Electrical Engineering
University of Colorado Denver**

Abstract

Graphics processors (GPUs) are designed to deliver large parallel computation and memory bandwidth. Traditionally, the architecture concepts of these processors grew as a response to the video game community which requires the processing and rendering of hundreds of millions of triangles per second - including calculations of lighting, geometry, color, and final pixel placement on the screen. The calculations of graphics are well suited for SIMD (Single Instruction Multiple Data) parallel processing, which is the computational model used in most graphics processors. Recently processor designers have found it increasingly difficult to provide performance improvements while managing power dissipation, chip temperature, and decreasing transistor reliability in traditional high-frequency monolithic processor-core designs. About five years ago, interest started to develop in harnessing the computational power of graphics processors for scientific computing. At that time, graphics processors were grossly inadequate for the task - due to several software barriers and a lack of scatter operations, poor latency, inadequate precision, and slow communication back to the CPU.

However, near the end of 2006, Nvidia released a new hardware/software programming model named CUDA (Compute Unified Device Architecture) which resolved most of these issues. CUDA is a highly scalable architecture based on the "C" language, which allows users to efficiently implement many types of scientific applications. Grid based calculations like finite difference computations are generally well suited for CUDA, as are some less regular problems such as computer vision. This seminar covers a short history of graphics and high-performance computing, and describes the fundamentals of the CUDA architecture. Several examples of real-world applications developed for GPUs will be examined. Computational performance speedups have been evaluated in the 20-100X range vs. a single CPU. The seminar will explore the potential benefits and issues of leveraging GPU systems for achieving energy-efficiency and high-performance in the domain of scientific applications.

Dan Connors is an Assistant Professor at the University of Colorado at Denver in the Department of Electrical Engineering. His research includes run-time compilation for energy control, temperature management, fault tolerance, and optimization of multi-core systems. He directs the DRACO research group which integrates compiler techniques, hardware performance monitors, and binary instrumentation systems in high-performance parallel computing. He received his Ph.D. in Computer Engineering from the University of Illinois at Urbana-Champaign in 2000.