UCF

Electrical and
Computer Engineering

Fall 2019 Seminar Series

Title

Exploiting Parallel Asynchronous Execution in AMD GPUs

Wednesday, November 6

4:00PM—5:00PM R1-101A

Abstract

Graphics Processing Units (GPUs) are the programmable accelerator of choice for a wide diversity of data-parallel problems. GPUs can process millions of polygons in parallel to provide rich visual experiences, as well as accelerate scientific compute workloads in the world's fastest supercomputers. While extremely powerful data-parallel processing engines, submitting large amounts of work to GPUs can be difficult. To combat this bottleneck, GPUs heavily rely on task partitioning and dynamic workload rebalancing to run efficiently, but often these features are not well-understood by programmers.

In this talk, we will split the presentation in two parts. First, we will provide an overview of AMD's latest GPU architecture, called RNDA. RDNA includes many features that improve GPU performance and energy efficiency including mechanisms for handling asynchronous tasks. Then in the second part, we will show one such scientific computing example, hyperplane sweep operations, that significant benefit from asynchronous execution. We will describe the data decomposition and scheduling techniques used to optimize the sweep operation, which achieve 41% speedup over the bulk synchronous implementation.

Brad Beckmann

Rex McCrary

Biography

Rex McCrary is an AMD Fellow and GPU Architect with management responsibilities for the GPU Subsystem Architecture Group. From a hardware perspective he focuses his team on designs that have the optimal balance of Performance, Power and Area. He believes the best innovation and the highest efficiencies can be gained by collaborating with different design groups across AMD. This focus can be traced back to his first degree at Carnegie-Mellon University, where interdisciplinary research is a fundamental principle. Here he received a BS in Electrical Engineering, Computer Engineering and Mathematics. His formal technology management training was received at the University of Miami, where he received a M.S. Management of Technology.

Brad Beckmann has been a member of AMD Research since 2007 and works in Bellevue, WA. Brad completed his PhD degree in the Department of Computer Science at the University of Wisconsin-Madison in 2006 where his doctoral research focused on physical and logical solutions to wire delay in CMP caches. While at AMD Research, he has worked on numerous projects related to memory consistency models, cache coherence, graphics, and on-chip networks. Currently, his primary research focuses on GPU compute solutions and broadening the impact of future AMD systems.