北京大学高能效计算与应用中心学术报告

Invited Talk, Center for Energy-Efficient Computing and Applications

SOFTWARE-HARDWARE INTEGRATED APPROACHES TO IMPROVING GPU RESOURCE UTILIZATION

Prof. Huiyang Zhou

Department of Electrical & Computer Engineering
North Carolina State University

2015年9月2日 星期三 10:00am

理科五号楼410会议室

ABSTRACT: Modern GPUs feature significant on-chip resources to support a high number of threads. Given various requirements from different applications, some resources can be severely underutilized while others become the performance bottleneck.

In this talk, we look into GPU resource management and introduce hardware and/or software approaches to effectively improving GPU resource utilization. In particular, we will showcase that coarse-grain resource management unnecessarily limits the realized thread-level parallelism. For shared memory, we analyze the lifetime of allocated shared memory and propose both software- and hardware-based ways to time-multiplex shared memory. For register files and warp schedulers, warp-level divergence is a fundamental reason to cause underutilization. To overcome this problem, we propose warp-level management for register files and warp schedulers. We also investigate compiler transformations to move data across different types of on-chip memory so as to balance the resource utilization and achieve performance portability.

BIOGRAPHY: Huiyang Zhou received the bachelor's degree in electrical engineering from Xian Jiaotong University, China, in 1992 and the Ph.D. degree in computer engineering from North Carolina State University in 2003. He is currently an associate professor in the Department of Electrical and Computer Engineering at North Carolina State University. Between 2003 and 2009, he was an assistant professor at the School of Electrical Engineering and Computer Science, University of Central Florida. His research focuses on high performance microarchitecture, low-power design, GPU Computing (General Purpose computing on Graphics Processing Units or GPGPU), architecture support for system dependability, and backend compiler optimization. He is a recipient of NSF CAREER award and a senior member of the ACM and IEEE.