NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

CS SEMINAR

Title: **Coordinated Static and Dynamic Cache Bypassing on GPUs**

Speaker: Yun (Eric) Liang
Assistant professor
School of EECS
Peking University

Date/Time: 18 August 2015, Tuesday, 11:00 AM to 12:30 PM

Venue: SR8, COM1-02-08

Chaired by: Dr Mitra, Tulika, Professor, School of Computing
(tulika@comp.nus.edu.sg)

Abstract:

The massive parallel architecture enables graphics processing units (GPUs) to boost performance for a wide range of applications. Recently, to broaden the scope of applications that can be accelerated by GPUs, GPU vendors have used caches in conjunction with scratchpad memory as on-chip memory in the new generations of GPUs. Unfortunately, GPU caches face many performance challenges that arise due to excessive thread contention for cache resource. Cache bypassing, where memory requests can selectively bypass the cache, is one solution that can help to mitigate the cache resource contention problem. In this work, we propose coordinated static and dynamic cache bypassing to improve application performance. At compile-time, we identify the global loads that indicate strong preferences for caching or bypassing through profiling. For the rest global loads, our dynamic cache bypassing has the flexibility to cache only a fraction of threads. Our coordinated static and dynamic cache bypassing technique achieves up to 2.28X (average 1.32X) performance speedup for a variety of GPU applications.

Biodata:

Yun (Eric) Liang is currently an assistant professor in School of EECS at Peking University, China. Before joining Peking University, he was a Research Scientist in Advanced Digital Science Center, University of Illinois at Champaign Urbana. He received the B.S degree from Tongji University, Shanghai, and the Ph.D degree in computer science from National University Singapore. He has published about 30 research papers in the top conferences and journals on compilation, computer architecture, and embedded system including HPCA, ISCA, DAC, CGO, ICCAD, FPGA, FCCM, etc. His work has received the Best Paper

Award of FCCM 2011 and Best Paper Award nominations from DAC 2012, FPT 2011, and CODES+ISSS 2008.