High-performance Memory Hierarchy Design for Throughput Processors

**ABSTRACT**

The continued growth of the computational capability of throughput processors has made throughput processors the platform of choice for a wide variety of high-performance computing applications. Graphics Processing Units (GPUs) are a prime example of throughput processors that can deliver high performance for applications ranging from typical graphics applications to general-purpose data parallel (GPGPU) applications. However, this success has been accompanied by new performance bottlenecks throughout the memory hierarchy of GPU-based systems. This talk identifies and eliminates performance bottlenecks caused by major sources of interference throughout the memory hierarchy. Specifically, we provide an in-depth analysis of bottlenecks at the caches, main memory, address translation and page-level transfers that significantly degrade the performance and efficiency of GPU-based systems. To minimize such performance bottlenecks, we introduce changes to the memory hierarchy for systems with GPUs that allow the entire memory hierarchy to be aware of applications' characteristics. Our proposal introduces a combination of GPU-aware cache and memory management techniques that effectively mitigates performance bottlenecks in the memory hierarchy of current and future GPU-based systems as well as other types of throughput processors.

**BIOGRAPHY**

Dr. Rachata Ausavarungnirun is a postdoctoral researcher at Carnegie Mellon University and a lecturer at the Sirindhorn International Thai-German Graduate School of Engineering at King Mongkut's University of Technology North Bangkok. His research spans multiple topics across computer architecture and system software with emphasis on GPU architecture, heterogeneous CPU-GPU architecture, management of GPUs in the cloud, memory subsystems, memory management, processing-in-memory, non-volatile memory, network-on-chip, and accelerator designs. He received his Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University in 2017.

All are welcome!

In case of questions, please contact Dr. Xue Chun Jason at Tel: 3442 9815, E-mail: jasonxue@cityu.edu.hk, or visit the CS Departmental Seminar Web at http://www.cs.cityu.edu.hk/