University of Pittsburgh | School of Computing and Information

Dissertation Defense
Doctor of Philosophy in Computer Science

"**Adaptive Memory Management for CPU-GPU Heterogeneous Systems**"
by **Debashis Ganguly**

Date: October 13, 2020
Time: 1:00 – 3:00 p.m.
Place: Virtual Defense

Committee:

- Rami Melhem, Department of Computer Science, School of Computing and Information
- Jun Yang, Department of Electrical and Computer Engineering, Swanson School of Engineering
- Youtao Zhang, Department of Computer Science, School of Computing and Information
- Bruce Childers, Department of Computer Science, School of Computing and Information

Abstract:
High compute-density with massive thread-level parallelism of Graphics Processing Units (GPUs) is behind their unprecedented adoption in systems ranging from data-centers to high performance computing installations. Currently, discrete GPU(s) combined with CPU via slow CPU-GPU interconnect dominate these computing platforms. The introduction of on-demand paging and fault-driven migration support in the newer generation GPUs, powered by software-managed unified memory runtime, simplified memory management in the CPU-GPU heterogeneous memory systems and ensured higher programmability. As GPUs are increasingly being used to accelerate general-purpose applications beyond traditional graphics processing, these systems raise a number of design challenges, including smart runtime systems, programming libraries, and micro-architecture.

One of the key challenges this dissertation aims to address is the performance slowdown under device memory oversubscription. When working set of the application exceeds device memory capacity, CPU-GPU interconnect traffic from page eviction and software prefetching becomes a major source of performance bottleneck. Firstly, this dissertation proposes a pre-eviction policy, that adapts the semantics of software prefetcher, to reduce the CPU-GPU interconnect traffic from unnecessary page thrashing. Secondly, this dissertation proposes an adaptive page migration and pinning strategy for the runtime that adapts to the irregularity in access pattern based on the frequency of memory access. Disparate applications demand special attention for memory management based on their workload characteristics, thread-level parallelism, and memory access pattern. Finally, this dissertation introduces a smart runtime that transparently caters to different classes of applications by unifying a wide array of memory management strategies. As GPUs are becoming integral part of commodity computing clusters, assuring system throughput and execution fairness is becoming a critical challenge for multi-tenant workloads. To this end, the dissertation proposes a CPU-GPU interconnect scheduler that provisions network traffic adapting to the disparate computation characteristics and bandwidth demands of participating applications in the composed

workload. By introducing all these techniques, the dissertation makes significant progress towards realizing the goal of developing an adaptive, smart software-managed runtime for CPU-GPU heterogeneous memory systems.