BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
Computer Science

COMPUTER SCIENCE RESEARCH SEMINAR

Neighborhood-aware address translation for irregular GPU applications

Seunghee Shin, Assistant Professor
Department of Computer Science, Binghamton University

Friday, October 5th at noon in room R15, Engineering Building

Abstract: Recent studies on commercial hardware demonstrated that irregular GPU workloads could bottleneck on virtual-to-physical address translations. GPU's single-.instruction multiple-thread (SIMT) execution can generate many concurrent memory accesses, all of which require address translation before accesses can complete. Unfortunately, many of these address translation requests often miss in the TLB, generating many concurrent page table walks. In this work, we investigate how to reduce address translation overheads for such applications. We observe that many of these concurrent page walk requests, while irregular from the perspective of a single GPU wavefront, still fall on neighboring virtual page addresses. The address mappings for these neighboring pages are typically stored in the same 64-byte cache line. Since cache lines are the smallest granularity of memory access, the page table walker implicitly reads address mappings (i.e., page table entries or PTEs) of many neighboring pages during the page walk of a single virtual address (VA). However, in the conventional hardware, mappings not associated with the original request are simply discarded. In this work, we propose mechanisms to coalesce the address translation needs of all pending page table walks in the same neighborhood that happens to have their address mappings fall on the same cache line. This is almost free; the page table walker (PTW) already reads a full cache line containing address mappings of all pages in the same neighborhood. We find this simple scheme can reduce the number of accesses to the in-memory page table by 37% on average. This speeds up a set of GPU workloads by an average of 1.7×.

Bio: Seunghee Shin received his Ph.D. degree from Electrical and Computer Engineering department at North Carolina State University, Raleigh, NC. His primary research interests lie in computer architecture and systems. Specifically, he has high interests in investigating the impact of emerging technologies on memory systems in modern processors. His research was published in the leading computer architecture venues such as the International Symposium on Computer Architecture (ISCA) or the International Symposium on Microarchitecture (MICRO). Besides, he has more than five years of professional system software development experiences in multiple companies where he engaged in mobile and storage system development projects. He also has M.S. degree in Computer Science from Northeastern University, MA, where he studied computer networks.

This event is funded by GSOCS, a subsidiary of GSO, using Student Activity Fee funds

Refreshments will be provided!