Computer Architecture Seminar

Efficient Memory Architecture Design for Emerging Technologies

Tuesday 11/13/2018
15:30~17:00 @ P711

Prof. Seunghee Shin
Binghamton University

Abstract

We are facing the end of Moore's law and Dennard scaling. Since transistor scaling is slowing down, the continued demand for power and performance efficient systems forces computer architects to re-think traditional computer architectures and search for alternative ways to realize better performance. In this trend, the computer architecture community pays attention to emerging technologies such as die-stacked DRAM, non-volatile memory, and heterogeneous architecture systems (HAS). However, since modern systems have been optimized for conventional CPUs and memories for decades, systems now need new designs in order to maximize their benefits from these new technologies.

Die-stacked DRAM technology enables a large Last Level Cache (LLC) that provides high bandwidth data access to the processor. However, it requires a large tag array that may take a significant portion of the on-chip SRAM budget. This overhead can be mitigated by adopting a large cache block size. However, the solution could waste memory bandwidth, also fetching unneeded bytes to LLC. To solve this problem, recent research relies on prediction technique, dividing the large block to multiple sub-blocks and only allocating sub-blocks that are predicted useful, while the remaining area of the large block is left empty, creating holes in the block. Unfortunately, holes create significant capacity overheads which could have been used for useful data. Meanwhile, byte-addressable non-volatile memory technology allows programmers to store important data in data structures in non-volatile main memory (NVMM) instead of serializing it to the file system, thereby providing a substantial performance boost. However, updating the data structures in main memory must be implemented in a way that ensures data consistency, so that software can recover to a consistent state in the event of system failures. This restriction demands new ordering constraints in the program, adding additional overheads. Lastly, the efforts to elevate GPUs to general purpose processors from graphics accelerators introduced HSA, a CPU-GPU integrated system sharing the same main memory. However, recent studies on commercial HSA hardware found that irregular GPU applications can bottleneck on virtual-to-physical address translations.

Bio

Seunghee Shin is currently an assistant professor in the Department of Computer Science at the State University of New York (SUNY) at Binghamton. He received his Ph.D. degree from Electrical and Computer Engineering department at North Carolina State University at Raleigh. He also has M.S. degree in Computer Science from Northeastern University, MA, where he studied computer networks. 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. 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.

Hosted by Gunjae Koo