Memory Systems in the Many-Core Era: Challenges, Opportunities, and Solution Directions

Onur Mutlu
Carnegie Mellon University
onur@cmu.edu
http://www.ece.cmu.edu/∼omutlu

The memory subsystem is a fundamental performance and energy bottleneck in almost all computing systems. Recent trends towards increasingly more cores on die, consolidation of diverse workloads on a single chip, and difficulty of DRAM scaling impose new requirements and exacerbate old demands on the memory system. In particular, the need for memory bandwidth and capacity is increasing, applications' interference in memory system increasingly limits system performance and makes the system hard to control, memory energy and power are key design concerns, and DRAM technology consumes significant amount of energy and does not scale down easily to smaller technology nodes. Fortunately, some promising solution directions exist.

In this talk, we will examine recent technology, application, and architecture trends motivating a fundamental rethinking of the memory hierarchy. Based on this motivation, we will describe requirements from an ideal memory system suitable for the many-core era. The talk will examine questions one would need to answer in approximating the ideal memory system and possible avenues that seem promising for the research community to explore. In particular, we will focus on the problem of uncontrolled inter-application interference in the memory system and draw upon our experiences in designing application-aware memory controllers and interconnects. We will make a case for application-aware design of memory systems and integrated/cooperative design of cores, interconnects, and memory components to optimize the overall system.

Brief Bio: Onur Mutlu is an Assistant Professor of ECE (and by courtesy CSD) at Carnegie Mellon University. His broader research interests are in computer architecture and systems, especially in the interactions between languages, operating systems, compilers, and microarchitecture. He enjoys teaching and researching important and relevant problems in computer architecture, including problems related to the design of memory systems, multi-core architectures, and scalable and efficient systems. He obtained his PhD and MS in ECE from the University of Texas at Austin (2006) and BS degrees in Computer Engineering and Psychology from the University of Michigan, Ann Arbor. His PhD dissertation was on efficient runahead execution to tolerate long main memory latencies. Prior to Carnegie Mellon, he worked at Microsoft Research (2006-2009), Intel Corporation, and Advanced Micro Devices. He was a recipient of the Microsoft Gold Star Award in 2008, ASPLOS Best Paper Award in 2010, NSF CAREER Award in 2010, University of Texas George H. Mitchell Award for Excellence in Graduate Research in 2005, and a number of "computer architecture top pick" paper selections by the IEEE Micro magazine over the past eight years. For more information, please see http://www.ece.cmu.edu/∼omutlu.

Categories and Subject Descriptors C.0 [Computer Systems Organization]: System architectures; C.1.2 [Computer Systems Organization]: Multiple Data Stream Architectures (Multiprocessors)

General Terms Algorithms, Design, Performance

Keywords Memory systems, multi-core, interconnects, parallelism