Keynote: Rethinking Memory System Design

Onur Mutlu

ETH Zürich and Carnegie Mellon University

Abstract

The memory system is a fundamental performance and energy bottleneck in almost all computing systems. Recent system design, application, and technology trends that require more capacity, bandwidth, efficiency, and predictability out of the memory system make it an even more important system bottleneck [36, 38]. At the same time, DRAM and flash technologies are experiencing difficult technology scaling challenges that make the maintenance and enhancement of their capacity, energy efficiency, and reliability significantly more costly with conventional techniques (see, for example [13, 17–19, 23, 25, 26, 29, 30, 40]). In fact, recent reliability issues with DRAM, such as the Row Hammer problem [23], are already threatening system security, predictability and robustness.

In this talk, we first discuss major challenges modern memory systems face in the presence of increasing demand for data and its fast analysis. We then examine some promising research and design directions to overcome these challenges and enable scalable memory systems for the future. We discuss three key design directions: 1) enabling new memory architectures, functions, interfaces, and better integration of memory and the rest of the system (e.g., [1, 2, 4, 12, 22, 27, 28, 32, 42–44]), 2) designing a memory system that intelligently employs emerging non-volatile memory (NVM) technologies (e.g., [24–26, 31, 33, 41, 49–51]), 3) reducing memory interference and providing predictable performance to applications sharing the memory system (e.g., [3, 14–16, 20–22, 35, 37, 39, 45–48]). If time permits, we will also touch upon our ongoing related work in combating scaling challenges of NAND flash memory (e.g., [5–11, 34]).

References