Architecting 3D Memory Systems: Different Constraints, Different Solutions

Dr. Moinuddin Qureshi, Georgia Institute of Technology
January 20, 2017
11:00 AM, 202 ECEC

Die-stacked 3D DRAM technology can provide low-energy high-bandwidth memory module by vertically integrating several dies within the same chip. However, the size of such 3D memory is unlikely to be sufficient to provide the full memory capacity, so future memory systems are likely to use 3D DRAM together with traditional off-chip DRAM. In-fact, such systems are already being announced by the industry. In this talk, I will discuss our work on architecting DRAM caches and share some of the insights and experiences that run counter to the well-established conventional wisdom in cache design.

I will show that some of the basic design decisions typically made for conventional caches (such as serialization of tag and data access, large associativity, and update of replacement state) are detrimental to the performance of DRAM caches, as they increase the hit latency. I will present *Alloy Cache*, a simple latency-optimized DRAM cache design that can outperform even an impractical SRAM tag-store design, which would incur an unacceptable overhead of several tens of megabytes. I will also present our *CAMEO* architecture that allows "Gigascale" DRAM caches to not only be transparently managed by the hardware but also to contribute to the OS-visible main-memory capacity. I will then analyze the bandwidth consumed by management operations (miss detection, install, write-backs etc.), show that these operations consume as much as 3x the cache bandwidth compared to the bandwidth consumed by data transfer on a cache hit, and present simple solutions to mitigate this bandwidth bloat. If time permits, we will also briefly discuss schemes that can proactively reduce the DRAM cache hit-rate to improve overall system performance.

Bio: Moinuddin Qureshi is an Associate Professor at the Georgia Institute of Technology. His research interests include computer architecture, scalable memory systems, fault tolerant computing, and analytical modeling of computer systems. Prior to joining Georgia Tech, he was a research staff member at IBM T.J. Watson Research Center, where he contributed to the design of efficient caching algorithms for Power 7 processors, and received the IBM outstanding technical achievement award for his studies on emerging memory technologies for server processors. He is an inventor of more than two-dozen U.S. patents (approved) and has 30+ publications in flagship architecture conferences. He served as the Program Chair for MICRO 2015 and is currently serving as the selection committee co-chair for IEEE MICRO Top Picks 2017. Dr. Qureshi received his Ph.D. (2007) and M.S. (2003), both in Electrical Engineering from the University of Texas at Austin, and B.E. (2000) degree from University of Mumbai.