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

Onur Mutlu

Carnegie Mellon University
http://www.ece.cmu.edu/∼omutlu

Abstract

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 [14], applications' interference in memory system increasingly limits system performance and makes the system hard to control [12], memory energy and power are key design concerns [8], and DRAM technology consumes significant amount of energy and does not scale down easily to smaller technology nodes [7]. 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 solving it by designing quality-of-service (QoS) aware memory controllers [5, 6, 9, 10, 11, 12], interconnects [1, 2, 13], and entire memory systems [3, 4]. We will make a case for application- and QoS-aware design of memory systems and integrated/cooperative design of cores, interconnects, and memory components to optimize the overall system.

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

General Terms Algorithms, Design, Performance

Keywords Memory systems, multi-core, chip multiprocessors, interconnects, quality of service

ISMM'11, June 4–5, 2011, San Jose, California, USA.
ACM 978-1-4503-0263-0/11/06.

References

[1] R. Das, O. Mutlu, T. Moscibroda, and C. Das. Application-aware prioritization mechanisms for on-chip networks. In *International Symposium on Microarchitecture (MICRO-42)*, 2009.

[2] R. Das, O. Mutlu, T. Moscibroda, and C. Das. Aergia: Exploiting packet latency slack in on-chip networks. In *International Symposium on Computer Architecture (ISCA-37)*, 2010.

[3] E. Ebrahimi, C. J. Lee, O. Mutlu, and Y. N. Patt. Fairness via source throttling: A configurable and high-performance fairness substrate for multi-core memory systems. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-XV)*, 2010.

[4] E. Ebrahimi, C. J. Lee, O. Mutlu, and Y. N. Patt. Prefetch-aware shared resource management for multi-core systems. In *International Symposium on Computer Architecture (ISCA-38)*, 2011.

[5] Y. Kim, D. Han, O. Mutlu, and M. Harchol-Balter. ATLAS: a scalable and high-performance scheduling algorithm for multiple memory controllers. In *International Symposium on High-Performance Computer Architecture (HPCA-16)*, 2010.

[6] Y. Kim, M. Papamichael, O. Mutlu, and M. Harchol-Balter. Thread cluster memory scheduling: Exploiting differences in memory access behavior. In *International Symposium on Microarchitecture (MICRO-43)*, 2010.

[7] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger. Architecting phase change memory as a scalable DRAM alternative. In *International Symposium on Computer Architecture (ISCA-36)*, 2009.

[8] C. Lefurgy, K. Rajamani, F. L. Rawson-III, W. M. Felter, M. Kistler, and T. W. Keller. Energy management for commercial servers. *IEEE Computer*, 36(12):39–48, 2003.

[9] T. Moscibroda and O. Mutlu. Memory performance attacks: Denial of memory service in multi-core systems. In *16th USENIX Security Symposium*, 2007.

[10] T. Moscibroda and O. Mutlu. Distributed order scheduling and its application to multi-core DRAM controllers. In *ACM Symposium on Principles of Distributed Computing (PODC-27)*, 2008.

[11] O. Mutlu and T. Moscibroda. Parallelism-aware batch scheduling: Enhancing both performance and fairness of shared DRAM systems. In *International Symposium on Computer Architecture (ISCA-35)*, 2008.

[12] O. Mutlu and T. Moscibroda. Stall-time fair memory access scheduling for chip multiprocessors. In *International Symposium on Microarchitecture (MICRO-40)*, 2007.

[13] G. Nychis, C. Fallin, T. Moscibroda, and O. Mutlu. Next generation on-chip networks: What kind of congestion control do we need? In *9th ACM Workshop on Hot Topics in Networks (HOTNETS)*, 2010.

[14] M. K. Qureshi, V. Srinivasan, and J. A. Rivers. Scalable high performance main memory system using phase-change memory technology. In *International Symposium on Computer Architecture (ISCA-36)*, 2009.