Keynote: Rethinking Memory System Design

Onur Mutlu
ETH Zürich

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 [48, 51]. 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 [17, 24–26, 28, 32, 34, 35, 38, 39, 54]). In fact, recent reliability issues with DRAM [46], such as the RowHammer problem [32], are already threatening system security and predictability.

In this talk, we first discuss major challenges facing modern memory systems in the presence of greatly increasing demand for data and its fast analysis. We then examine some promising research and design directions to overcome these challenges and thus enable scalable memory systems for the future. We discuss three key solution directions: 1) enabling new memory architectures, functions, interfaces, and better integration of memory and the rest of the system (e.g., [1, 2, 4, 13–15, 20–23, 25, 31, 36–38, 41, 52–54, 56–58]), 2) designing a memory system that intelligently employs emerging non-volatile memory (NVM) technologies and coordinates memory and storage management (e.g., [33–35, 40, 44, 55, 63–65]), 3) reducing memory interference and providing predictable performance to applications sharing the memory system (e.g., [3, 16, 18, 19, 27, 29, 30, 47, 49, 50, 59–62]). If time permits, we will also touch upon our ongoing related work in combating scaling challenges of NAND flash memory (e.g., [5–12, 42, 43, 45]).

References

[1] J. Ahn et al. PIM-Enabled Instructions: A Low-Overhead, Locality-Aware Processing-in-Memory Architecture. In *ISCA*, 2015.
[2] J. Ahn et al. A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing. In *ISCA*, 2015.
[3] R. Ausavarungnirun et al. Staged memory scheduling: Achieving high performance and scalability in heterogeneous systems. In *ISCA*, 2012.
[4] A. Boroumand et al. LazyPIM: An Efficient Cache Coherence Mechanism for Processing-in-Memory. *CAL*, 2016.
[5] Y. Cai et al. Error patterns in MLC NAND flash memory: Measurement, characterization, and analysis. In *DATE*, 2012.
[6] Y. Cai et al. Flash Correct-and-Refresh: Retention-aware error management for increased flash memory lifetime. In *ICCD*, 2012.
[7] Y. Cai et al. Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis and modeling. In *DATE*, 2013.
[8] Y. Cai et al. Program interference in MLC NAND flash memory: Characterization, modeling, and mitigation. In *ICCD*, 2013.
[9] Y. Cai et al. Error Analysis and Retention-Aware Error Management for NAND Flash Memory. *ITJ*, 2013.
[10] Y. Cai et al. Neighbor-cell assisted error correction for MLC NAND flash memories. In *SIGMETRICS*, 2014.
[11] Y. Cai et al. Read Disturb Errors in MLC NAND Flash Memory: Characterization, Mitigation, and Recovery. In *DSN*, 2015.
[12] Y. Cai et al. Data retention in MLC NAND flash memory: Characterization, optimization and recovery. In *HPCA*, 2015.
[13] K. Chang et al. Improving DRAM performance by parallelizing refreshes with accesses. In *HPCA*, 2014.
[14] K. Chang et al. Understanding Latency Variation in Modern DRAM Chips: Experimental Characterization, Analysis, and Optimization. *SIGMETRICS*, 2016.
[15] K. Chang et al. Low-Cost Inter-Linked Subarrays (LISA): Enabling Fast Inter-Subarray Data Movement in DRAM. In *HPCA*, 2016.
[16] R. Das et al. Application-to-core mapping policies to reduce memory system interference in multi-core systems. In *HPCA*, 2013.
[17] H. David et al. Memory power management via dynamic voltage/frequency scaling. In *ICAC*, 2011.
[18] E. Ebrahimi et al. Fairness via source throttling: a configurable and high-performance fairness substrate for multi-core memory systems. *ASPLOS*, 2010.
[19] E. Ebrahimi et al. Prefetch-aware shared-resource management for multi-core systems. In *ISCA*, 2011.
[20] M. Hashemi et al. Accelerating Dependent Cache Misses with an Enhanced Memory Controller. In *ISCA*, 2016.
[21] H. Hassan et al. ChargeCache: Reducing DRAM Latency by Exploiting Row Access Locality. In *HPCA*, 2016.
[22] K. Hsieh et al. Accelerating Pointer Chasing in 3D-Stacked Memory: Challenges, Mechanisms, Evaluation. *ICCD*, 2016.
[23] K. Hsieh et al. Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems. *ISCA*, 2016.
[24] U. Kang et al. Co-architecting controllers and DRAM to enhance DRAM process scaling. In *The Memory Forum*, 2014.
[25] S. Khan et al. The efficacy of error mitigation techniques for DRAM retention failures: A comparative experimental study. In *SIGMETRICS*, 2014.
[26] S. Khan et al. PARBOR: An Efficient System-Level Technique to Detect Data-Dependent Failures in DRAM. In *DSN*, 2016.
[27] H. Kim et al. Bounding memory interference delay in COTS-based multi-core systems. In *RTAS*, 2014.
[28] Y. Kim. Ramulator: A Fast and Extensible DRAM Simulator. *CAL*, 2015.
[29] Y. Kim et al. ATLAS: a scalable and high-performance scheduling algorithm for multiple memory controllers. In *HPCA*, 2010.
[30] Y. Kim et al. Thread cluster memory scheduling: Exploiting differences in memory access behavior. In *MICRO*, 2010.
[31] Y. Kim et al. A case for subarray-level parallelism (SALP) in DRAM. In *ISCA*, 2012.
[32] Y. Kim et al. Flipping bits in memory without accessing them: An experimental study of DRAM disturbance errors. In *ISCA*, 2014.
[33] E. Kultursay et al. Evaluating STT-RAM as an energy-efficient main memory alternative. In *ISPASS*, 2013.
[34] B. C. Lee et al. Architecting phase change memory as a scalable DRAM alternative. In *ISCA*, 2009.
[35] B. C. Lee et al. Phase change memory architecture and the quest for scalability. *CACM*, 2010.
[36] D. Lee et al. Tiered-latency DRAM: A low latency and low cost DRAM architecture. In *HPCA*, 2013.
[37] D. Lee et al. Adaptive-latency DRAM: Optimizing DRAM timing for the common-case. In *HPCA*, 2015.
[38] J. Liu et al. RAIDR: Retention-aware intelligent DRAM refresh. *ISCA*, 2012.
[39] J. Liu et al. An experimental study of data retention behavior in modern DRAM devices: Implications for retention time profiling mechanisms. *ISCA*, 2013.
[40] Y. Lu et al. Loose-ordering consistency for persistent memory. *ICCD*, 2014.
[41] Y. Luo et al. Characterizing application memory error vulnerability to optimize data center cost via heterogeneous-reliability memory. In *DSN*, 2014.
[42] Y. Luo et al. WARM: Improving NAND Flash Memory Lifetime with Write-hotness Aware Retention Management. *MSST*, 2015.
[43] Y. Luo et al. Enabling Accurate and Practical Online Flash Channel Modeling for Modern MLC NAND Flash Memory. *JSAC*, 2016.
[44] J. Meza et al. A case for efficient hardware-software cooperative management of storage and memory. In *WEED*, 2013.
[45] J. Meza et al. A Large-Scale Study of Flash Memory Errors in the Field. In *SIGMETRICS*, 2015.
[46] J. Meza et al. Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field. In *DSN*, 2015.
[47] S. Muralidhara et al. Reducing memory interference in multi-core systems via application-aware memory channel partitioning. In *MICRO*, 2011.
[48] O. Mutlu. Memory scaling: A systems architecture perspective. *IMW*, 2013.
[49] O. Mutlu and T. Moscibroda. Stall-time fair memory access scheduling for chip multiprocessors. In *MICRO*, 2007.
[50] O. Mutlu and T. Moscibroda. Parallelism-aware batch scheduling: Enhancing both performance and fairness of shared DRAM systems. In *ISCA*, 2008.
[51] O. Mutlu and L. Subramanian. Research problems and opportunities in memory systems. *SUPERFRI*, 2014.
[52] A. Pattnaik et al. Scheduling Techniques for GPU Architectures with Processing-In-Memory Capabilities. *PACT*, 2016.
[53] G. Pekhimenko et al. Linearly Compressed Pages: A Low-Complexity, Low-Latency Main Memory Compression Framework. *MICRO*, 2013.
[54] M. K. Qureshi et al. AVATAR: A Variable-Retention-Time (VRT) Aware Refresh for DRAM Systems. In *DSN*, 2015.
[55] J. Ren et al. ThyNVM: Enabling Software-Transparent Crash Consistency in Persistent Memory Systems. In *MICRO*, 2015.
[56] V. Seshadri et al. RowClone: Fast and efficient In-DRAM copy and initialization of bulk data. In *MICRO*, 2013.
[57] V. Seshadri et al. Gather-Scatter DRAM: In-DRAM Address Translation to Improve the Spatial Locality of Non-unit Strided Accesses. In *MICRO*, 2015.
[58] V. Seshadri et al. Fast Bulk Bitwise AND and OR in DRAM. *CAL*, 2015.
[59] L. Subramanian et al. MISE: Providing performance predictability and improving fairness in shared main memory systems. In *HPCA*, 2013.
[60] L. Subramanian et al. The blacklisting memory scheduler: Achieving high performance and fairness at low cost. In *ICCD*, 2014.
[61] L. Subramanian et al. The application slowdown model: Quantifying and controlling the impact of inter-application interference at shared caches and main memory. In *MICRO*, 2015.
[62] H. Usui et al. DASH: Deadline-Aware High-Performance Memory Scheduler for Heterogeneous Systems with Hardware Accelerators. *TACO*, 2016.
[63] H. Yoon et al. Row buffer locality aware caching policies for hybrid memories. In *ICCD*, 2012.
[64] H. Yoon et al. Efficient data mapping and buffering techniques for multi-level cell phase-change memories. *TACO*, 2014.
[65] J. Zhao et al. FIRM: Fair and high-performance memory control for persistent memory systems. In *MICRO*, 2014.