Retrospective: An Experimental Study of Data Retention Behavior in Modern DRAM Devices: Implications for Retention Time Profiling Mechanisms

Onur Mutlu
ETH Zürich

***Abstract*—DRAM is the prevalent main memory technology used in almost all computers. Data is represented as charge in a DRAM cell. Unfortunately, a DRAM cell loses its stored charge over time and thus needs to be refreshed periodically. How often a cell needs to be refreshed depends on its minimum data retention time, which is dependent on various factors. Accurately identifying the minimum data retention time of each DRAM cell is necessary to 1) correctly determine the minimum refresh rate of a DRAM chip to maintain data integrity, and 2) enable techniques that eliminate unnecessary refresh operations by refreshing each DRAM row at the minimum refresh rate it needs for reliable operation.**

Our ISCA 2013 paper [1] provides a fundamental empirical understanding of two major factors that make it very difficult to determine the minimum data retention time of a DRAM cell, based on the first comprehensive experimental characterization of retention time behavior of a large number of modern commodity DRAM chips from 5 major vendors. We study the prevalence, effects, and technology scaling characteristics of two significant phenomena: 1) *data pattern dependence (DPD)*, where the minimum retention time of a DRAM cell is affected by data stored in other DRAM cells, and 2) *variable retention time (VRT)*, where the minimum retention time of a DRAM cell changes unpredictably over time. To this end, we built a flexible FPGA-based testing infrastructure to test DRAM chips, which has enabled a large amount of further experimental research in DRAM. Our ISCA 2013 paper's results using this infrastructure clearly demonstrate that DPD and VRT phenomena are significant issues that must be addressed for correct operation in DRAM-based systems and their effects are getting worse as DRAM scales to smaller technology node sizes. Our work also provides ideas on how to accurately identify data retention times in the presence of DPD and VRT, e.g., online profiling with error correcting codes, which later works examined and enabled. Most modern DRAM chips now incorporate ECC, especially to account for VRT effects [2].

This short retrospective provides a brief analysis of our ISCA 2013 paper and its impact. We describe why we did the work, what we found and its implications, what the findings as well as the infrastructure we built to discover them have enabled in later works, and our thoughts on what the future may bring.

I. Background

My group has been working on the DRAM refresh problem since 2010 and our major work RAIDR [3] was published at ISCA 2012. Our goal in RAIDR was to eliminate unnecessary refresh operations at low cost by refreshing each DRAM row only as frequently as required by the minimum data retention time of the row. As described in a separate retrospective in this issue, our RAIDR work demonstrated large performance improvements and energy savings with a simple memory controller based implementation. However, we were not satisfied with the simplistic retention time profiling mechanism assumed in RAIDR (which was also assumed in other prior works). RAIDR relied on accurate identification of the minimum data retention time of every DRAM cell, which we thought was a difficult task. We wanted to make such retention time profiling practical. So, we set out to rigorously understand the difficulty of DRAM data retention time identification using an empirical approach. No prior work at the time provided real data on the retention characteristics of state-of-the-art DRAM chips, let alone a detailed empirical analysis of major problems that make retention time profiling challenging and how DRAM technology scaling affects those challenges. In fact, no infrastructure to study these characteristics existed (or was available to us). We decided to build our own FPGA-based infrastructure to characterize real DRAM chips in a flexible manner so that we could change the refresh rate, data patterns, and other major parameters. This infrastructure, which took us more than a year to build and which we later open sourced as SoftMC [4, 5] and DRAM Bender [6, 7], enabled us (and others) to empirically study and understand many interesting characteristics of modern DRAM chips over the course of more than a decade.

Our ISCA 2013 paper is a product of this goal and effort. Our work was generously supported especially by the Samsung DRAM Design Team and Intel Memory Architecture Labs, both technically and funding-wise. With close technical support from Intel (especially Chris Wilkerson, who is a co-author), we built our FPGA-based DDR3 DRAM testing infrastructure. Two of my students (also co-authors) and I spent part of the summer of 2012 at Intel to work closely with our collaborators. During this timeframe, we finalized the calibration and stabilization of our infrastructure. We performed many experiments to study both well-known properties of retention time characteristics (e.g., temperature dependence) as well as less well studied characteristics (e.g., DPD and VRT phenomena) of modern DRAM chips at a scale that was not reported before. We were especially interested in empirically understanding how technology scaling affected such characteristics, since it was clear that data retention and thus refresh was a major technology scaling challenge in DRAM, as indicated by prior and later works (e.g., [2, 8, 9]).

II. Contributions and Impact

Our paper is the first to comprehensively examine data retention time behavior of modern DRAM chips, uncovering real data and insights on two major phenomena that make retention time identification extremely challenging. Prior works were limited to simulation or had very small sample sizes, and almost none of them examined modern DDR3 DRAM chips or technology scaling. Many device- or circuit-level works did not study DPD or VRT. No architecture- and system-level work to reduce refresh overhead discussed DPD or VRT. Our work enabled a new understanding and demonstrated the true difficulty of a major problem in DRAM technology scaling, by providing valuable data that was available nowhere else (at least publicly).

Our key results demonstrate that data retention times of modern DRAM chips are indisputably getting worse in newer-generation DRAM chips, indicating that refresh is becoming a larger problem with technology scaling. Ditto for DPD and VRT. For example, we showed that 1) the retention failure coverage of a given data pattern becomes smaller for newer-generation DRAM chips, 2) VRT is a widespread phenomenon in modern DRAM devices, causing significant dynamic changes in minimum retention time. These were the first results of their kind.

Our results indicated that many prior proposals (e.g., [3, 10–14]) that rely on accurate retention time identification to eliminate refreshes would not work reliably as they do not take into account DPD or VRT. They also put into question whether existing refresh rates are enough to guarantee error-free operation in DRAM chips being used in the field (especially in the presence of VRT). As DRAM technology scales, would it be easy to accurately determine retention times to ensure data integrity even if we maintained a conservative refresh rate for all DRAM cells?

Based on the understanding we developed, we proposed ideas and avenues for future work on how to tackle the DPD and VRT problems ($5.2 & $6.3 in [1]). We advocated the use of ECC in DRAM chips to detect and/or correct any retention errors that might not be identified after rigorous testing (offline or online). Most modern DRAM chips now incorporate ECC (see [15–17]), especially to account for VRT effects [2]. We also advocated the use of online profiling together with ECC to enable reliable identification of retention times, an approach later works

rigorously investigated and enabled (e.g., [16–23]). As such, our ISCA 2013 paper enabled system-level techniques to overcome a major DRAM scaling challenge, an approach we call *system-DRAM co-design* [24, 25]. We believe developing such system-level techniques that can detect and exploit DRAM characteristics online, during system operation, will be increasingly valuable as such characteristics will become much more difficult to accurately determine and exploit due to technology scaling.

A key contribution of our work was the development of our flexible FPGA-based DRAM testing infrastructure, which was the first of its kind. It enabled a large amount of research into DRAM chips by enabling rigorous experimental study of real DRAM chip characteristics, including the rigorous study of the RowHammer vulnerability [6, 26–36], another major DRAM technology scaling problem. We discuss some new insights and studies enabled by this infrastructure in our RAIDR retrospective and our SoftMC [4] and DRAM Bender [6] works.

III. Influence on Later Works

Many later works (e.g., [16–23]) ensued to solve the DPD & VRT problems. Some provided a more detailed characterization of the DPD and VRT phenomena: [18] analyzed both DPD & VRT and examined the effectiveness of online profiling versus ECC of varying strength. AVATAR [19] provided heterogeneous refresh rates using a combination of online profiling, ECC, and memory scrubbing, working from the empirical observation that new VRT errors are discovered infrequently at a steady rate. PARBOR [20] introduced detailed DPD analyses and a new technique to efficiently detect data-dependent failures. REAPER [23] analyzed the DPD & VRT phenomena in newer LPDDR4 DRAM chips, demonstrating that the problems are getting worse, and developed the reach profiling technique to tolerate the two problems. We believe AVATAR & REAPER enabled practical ways of exploiting heterogeneous retention times.

ECC is mainstream in DRAM chips today [15–17]. We believe this is a direct result of the analysis that showed the prevalence and importance of VRT and the difficulty of handling VRT-caused retention errors due to their fundamentally unpredictable nature. A later work by Samsung & Intel engineers [2] described that ECC is needed to deal with VRT, just as our ISCA'13 paper advocated.

IV. Summary and Future Outlook

Our ISCA 2013 paper was a nice example of harmonious collaboration between academia and industry: Intel helped us build the infrastructure and both Intel & Samsung gave us significant technical feedback along with generous funding. Our paper also highlights the importance of investing into building infrastructure to analyze real chips: doing so enabled not only the new understanding developed in our work, but also many future works that analyzed various other DRAM characteristics (e.g., [6, 16–23, 26–42]) and uncovered fascinating undocumented capabilities in real DRAM chips, e.g., the ability to perform data copy/initialization and bitwise operations [43–47], implement physical unclonable functions [48], and generate true random numbers [49, 50].

Since ISCA 2013, we have come a long way in understanding fundamental characteristics of DRAM devices and enabling practical solutions to overcome DRAM shortcomings. Yet, there is a lot more to be empirically discovered and understood in DRAM to solve the fundamental scaling, performance, and energy challenges of the technology (as shown by very recent works in 2022-2023, e.g. [6, 30–33, 45, 47]), which can enable solutions also applicable to other technologies. We conclude that the future is bright in experimental memory systems research using real memory chips.

References

[1] J. Liu *et al.*, "An Experimental Study of Data Retention Behavior in Modern DRAM Devices: Implications for Retention Time Profiling Mechanisms," *ISCA*, 2013.
[2] U. Kang *et al.*, "Co-Architecting Controllers and DRAM to Enhance DRAM Process Scaling," in *The Memory Forum*, 2014.
[3] J. Liu *et al.*, "RAIDR: Retention-Aware Intelligent DRAM Refresh," in *ISCA*, 2012.
[4] H. Hassan *et al.*, "SoftMC: A Flexible and Practical Open-Source Infrastructure for Enabling Experimental DRAM Studies," in *HPCA*, 2017.
[5] SoftMC Source Code, https://github.com/CMU-SAFARI/SoftMC.
[6] A. Olgun *et al.*, "DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips," *TCAD*, 2023.
[7] "DRAM Bender," https://github.com/CMU-SAFARI/DRAM-Bender.
[8] J. A. Mandelman *et al.*, "Challenges and Future Directions for the Scaling of Dynamic Random-Access Memory (DRAM)," *IBM JRD*, 2002.
[9] W. Kim, "A 1.1V 16Gb DDR5 DRAM with Probabilistic-Aggressor Tracking, Refresh-Management Functionality, Per-Row Hammer Tracking, a Multi-Step Precharge, and Core-Bias Modulation for Security and Reliability Enhancement," in *ISSCC*, 2023.
[10] J.-H. Ahn *et al.*, "Adaptive self refresh scheme for battery operated high-density mobile DRAM applications," in *ASSCC*, 2006.
[11] T. Ohsawa *et al.*, "Optimizing the DRAM Refresh Count for Merged DRAM/Logic LSIs," in *ISLPED*, 1998.
[12] J. Kim and M. C. Papaefthymiou, "Dynamic memory design for low data-retention power," in *PATMOS*, 2000.
[13] R. Venkatesan *et al.*, "Retention-Aware Placement in DRAM (RAPID): Software Methods for Quasi-Non-Volatile DRAM," in *HPCA*, 2006.
[14] K. Yanagisawa, "Semiconductor Memory," 1988, U.S. Patent 4,736,344.
[15] M. Patel *et al.*, "Understanding and Modeling On-die Error Correction in Modern DRAM: An Experimental Study using Real Devices," in *DSN*, 2019.
[16] M. Patel *et al.*, "Bit-Exact ECC Recovery (BEER): Determining DRAM On-Die ECC Functions by Exploiting DRAM Data Retention Characteristics," in *MICRO*, 2020.
[17] M. Patel *et al.*, "HARP: Practically and Effectively Identifying Uncorrectable Errors in Main Memory Chips That Use On-Die ECC," in *MICRO*, 2021.
[18] S. Khan, "The Efficacy of Error Mitigation Techniques for DRAM Retention Failures: A Comparative Experimental Study," in *SIGMETRICS*, 2014.
[19] M. K. Qureshi *et al.*, "AVATAR: A Variable-Retention-Time (VRT) Aware Refresh for DRAM Systems," in *DSN*, 2015.
[20] S. Khan *et al.*, "PARBOR: An Efficient System-Level Technique to Detect Data Dependent Failures in DRAM," in *DSN*, 2016.
[21] S. Khan *et al.*, "A Case for Memory Content-Based Detection and Mitigation of Data-Dependent Failures in DRAM," *CAL*, 2016.
[22] S. Khan *et al.*, "Detecting and Mitigating Data-Dependent DRAM Failures by Exploiting Current Memory Content," in *MICRO*, 2017.
[23] M. Patel, "The Reach Profiler (REAPER): Enabling the Mitigation of DRAM Retention Failures via Profiling at Aggressive Conditions," in *ISCA*, 2017.
[24] O. Mutlu, "Memory Scaling: A Systems Architecture Perspective," *IMW*, 2013.
[25] O. Mutlu and L. Subramanian, "Research Problems and Opportunities in Memory Systems," *SUPERFRI*, 2014.
[26] Y. Kim *et al.*, "Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors," in *ISCA*, 2014.
[27] J. S. Kim *et al.*, "Revisiting RowHammer: An Experimental Analysis of Modern DRAM Devices and Mitigation Techniques," in *ISCA*, 2020.
[28] P. Frigo *et al.*, "TRRespass: Exploiting the Many Sides of Target Row Refresh," in *S&P*, 2020.
[29] H. Hassan *et al.*, "Uncovering In-DRAM RowHammer Protection Mechanisms: A New Methodology, Custom RowHammer Patterns, and Implications," in *MICRO*, 2021.
[30] A. G. Yağlikci *et al.*, "HiRA: Hidden Row Activation for Reducing Refresh Latency of Off-the-Shelf DRAM Chips," in *MICRO*, 2022.
[31] A. G. Yağlıkcı *et al.*, "Understanding RowHammer Under Reduced Wordline Voltage: An Experimental Study Using Real DRAM Devices," in *DSN*, 2022.
[32] A. Olgun *et al.*, "An Experimental Analysis of RowHammer in HBM2 DRAM Chips," in *DSN Disrupt*, 2023.
[33] H. Luo *et al.*, "RowPress: Amplifying Read Disturbance in Modern DRAM Chips," in *ISCA*, 2023.
[34] M. Farmani *et al.*, "RHAT: Efficient RowHammer-Aware Test for Modern DRAM Modules," in *ETS*, 2021.
[35] O. Mutlu and J. S. Kim, "RowHammer: A Retrospective," *IEEE TCAD*, 2019.
[36] O. Mutlu *et al.*, "Fundamentally Understanding and Solving RowHammer," in *ASP-DAC*, 2023.
[37] D. Lee *et al.*, "Adaptive-Latency DRAM: Optimizing DRAM Timing for the Common-Case," in *HPCA*, 2015.
[38] K. K. Chang *et al.*, "Understanding Latency Variation in Modern DRAM Chips: Experimental Characterization, Analysis, and Optimization," in *SIGMETRICS*, 2016.
[39] D. Lee *et al.*, "Design-induced Latency Variation in Modern DRAM Chips: Characterization, Analysis, and Latency Reduction Mechanisms," *POMACS*, 2017.
[40] J. Kim *et al.*, "Solar-DRAM: Reducing DRAM Access Latency by Exploiting the Variation in Local Bitlines," in *ICCD*, 2018.
[41] K. Chang *et al.*, "Understanding Reduced-Voltage Operation in Modern DRAM Devices: Experimental Characterization, Analysis, and Mechanisms," in *SIGMETRICS*, 2017.
[42] S. Ghose *et al.*, "What Your DRAM Power Models Are Not Telling You: Lessons from a Detailed Experimental Study," in *SIGMETRICS*, 2018.
[43] V. Seshadri *et al.*, "Fast Bulk Bitwise AND and OR in DRAM," *CAL*, 2015.
[44] V. Seshadri *et al.*, "Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology," in *MICRO*, 2017.
[45] A. Olgun *et al.*, "PiDRAM: A Holistic End-to-end FPGA-based Framework for Processing-in-DRAM"," *TACO*, 2023.
[46] F. Gao *et al.*, "ComputeDRAM: In-Memory Compute Using Off-the-Shelf DRAMs," in *MICRO*, 2019.
[47] F. Gao *et al.*, "FracDRAM: Fractional Values in Off-the-Shelf DRAM," in *MICRO*, 2022.
[48] J. S. Kim *et al.*, "The DRAM Latency PUF: Quickly Evaluating Physical Unclonable Functions by Exploiting the Latency-Reliability Tradeoff in Modern Commodity DRAM Devices," in *HPCA*, 2018.
[49] J. Kim, "D-RaNGe: Using Commodity DRAM Devices to Generate True Random Numbers with Low Latency and High Throughput," in *HPCA*, 2019.
[50] A. Olgun *et al.*, "QUAC-TRNG: High-Throughput True Random Number Generation Using Quadruple Row Activation in Commodity DRAMs," in *ISCA*, 2021.