Understanding Reduced-Voltage Operation in Modern DRAM Devices: Experimental Characterization, Analysis, and Mechanisms*

Kevin K. Chang[†] Abdullah Giray Yağlıkçı[†] Saugata Ghose[†] Aditya Agrawal[¶] Niladrish Chatterjee[¶]
Abhijith Kashyap[¶] Donghyuk Lee[¶] Mike O'Connor[¶,‡] Hasan Hassan[§] Onur Mutlu[§,†]
[†]Carnegie Mellon University [¶]NVIDIA [‡]The University of Texas at Austin [§]ETH Zürich

ABSTRACT

The energy consumption of DRAM is a critical concern in modern computing systems. Improvements in manufacturing process technology have allowed DRAM vendors to lower the DRAM supply voltage conservatively, which reduces some of the DRAM energy consumption. We would like to reduce the DRAM supply voltage more aggressively, to further reduce energy. Aggressive supply voltage reduction requires a thorough understanding of the effect voltage scaling has on DRAM access latency and DRAM reliability.

In this paper, we take a comprehensive approach to understanding and exploiting the latency and reliability characteristics of modern DRAM when the supply voltage is lowered below the nominal voltage level specified by manufacturers. Using an FPGA-based testing platform, we perform an experimental study of 124 real DDR3L (low-voltage) DRAM chips manufactured recently by three major DRAM vendors. Our extensive experimental characterization yields four major observations on how DRAM latency, reliability, and data retention are affected by reduced voltage.

First, we observe that we can reliably access data when DRAM supply voltage is lowered below the nominal voltage level, *until a certain voltage value*, V_{min}, which is the minimum voltage level at which no bit errors occur. Furthermore, we find that we can reduce the voltage below V_{min} to attain further energy savings, but that errors start occurring in some of the data read from memory. As we drop the voltage further below V_{min}, the number of erroneous bits of data increases exponentially.

Second, we observe that while reducing the voltage below V_{min} introduces bit errors in the data, we can prevent these errors if we increase the latency of three major DRAM operations, i.e., activation, restoration, and precharge. When the supply voltage is reduced, the DRAM cell capacitor charge takes a longer time to change, thereby causing these DRAM operations to become slower to complete. Errors are introduced into the data when the memory controller does *not* account for this slowdown in the DRAM operations. We find that if the memory controller allocates extra time for these operations to finish when the supply voltage is below V_{min}, errors no longer occur. We validate, analyze, and explain this behavior using detailed circuit-level simulations.

Third, we observe that when only a small number of errors occur due to reduced supply voltage, these errors tend to *cluster* physically in certain *regions* of a DRAM chip, as opposed to being randomly distributed throughout the chip. This observation implies that when we reduce the supply voltage to the DRAM array, we need to increase the fundamental operation latencies for *only* the regions where errors can occur.

Fourth, we observe that reducing the supply voltage does *not* affect the data retention guarantees of DRAM. Commodity DRAM chips guarantee that all cells can safely retain data for 64ms, after which the cells are *refreshed* to replenish charge that leaks out of the capacitors. Even when we reduce the supply voltage, the rate at which charge leaks from the capacitors is so slow that no data is lost during the 64ms refresh interval at both 20℃ and 70℃.

Based on our observations, we propose a new DRAM energy reduction mechanism, called *Voltron*. The key idea of Voltron is to use a performance model to determine by how much we can reduce the supply voltage without introducing errors and without exceeding a user-specified threshold for performance loss. Unlike prior works, Voltron does *not* reduce the voltage of the *peripheral circuitry*, which is responsible for transferring commands and data between the memory controller and the DRAM chip. If Voltron were to reduce the voltage of the peripheral circuitry, we would have to reduce the operating frequency of DRAM. A reduction in the operating frequency reduces the memory data throughput, which can significantly degrade the performance of applications that require high memory bandwidth. Our evaluations show that Voltron reduces the average DRAM and system energy consumption by 10.5% and 7.3%, respectively, while limiting the average system performance loss to only 1.8%, for a variety of memory-intensive quad-core workloads. We also show that Voltron significantly outperforms prior dynamic voltage and frequency scaling mechanisms for DRAM.

KEYWORDS

DRAM; Voltage Reduction; Memory Latency; Reliability; Performance; Energy; DRAM Characterization; Memory Systems

ACM Reference format:
Kevin K. Chang, Abdullah Giray Yağlıkçı, Saugata Ghose, Aditya Agrawal, Niladrish Chatterjee, Abhijith Kashyap, Donghyuk Lee, Mike O'Connor, Hasan Hassan, and Onur Mutlu. 2017. Understanding Reduced-Voltage Operation in Modern DRAM Devices: Experimental Characterization, Analysis, and Mechanisms. In *Proceedings of SIGMETRICS '17, June 5–9, 2017, Urbana-Champaign, IL, USA,* , 1 pages.
DOI: http://dx.doi.org/10.1145/3078505.3078590

*The full version of the paper is available at http://www.ece.cmu.edu/~safari/pubs.html

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGMETRICS '17, June 5–9, 2017, Urbana-Champaign, IL, USA
© 2017 Copyright held by the owner/author(s). ACM ISBN 978-1-4503-5032-7/17/06.
DOI: http://dx.doi.org/10.1145/3078505.3078590