# Design-Induced Latency Variation in Modern DRAM Chips: Characterization, Analysis, and Latency Reduction Mechanisms

Extended Abstract*

Donghyuk Lee[†‡] Samira Khan[ᵡ] Lavanya Subramanian[†] Saugata Ghose[†]
Rachata Ausavarungnirun[†] Gennady Pekhimenko[†¶] Vivek Seshadri[†¶] Onur Mutlu[†§]

[†]*Carnegie Mellon University* [‡]*NVIDIA* [ᵡ]*University of Virginia* [¶]*Microsoft Research* [§]*ETH Zürich*

## ABSTRACT

Variation has been shown to exist across the cells within a modern DRAM chip. Prior work has studied and exploited several forms of variation, such as manufacturing-process- or temperature-induced variation. We empirically demonstrate a new form of variation that exists within a real DRAM chip, *induced by the design and placement* of different components in the DRAM chip: different regions in DRAM, based on their relative distances from the peripheral structures, require different minimum access latencies for reliable operation. In particular, we show that in most real DRAM chips, cells closer to the peripheral structures can be accessed much faster than cells that are farther. We call this phenomenon *design-induced variation in DRAM*. Our goals are to *i)* understand design-induced variation that exists in real, state-of-the-art DRAM chips, *ii)* exploit it to develop low-cost mechanisms that can dynamically find and use the *lowest latency at which to operate a DRAM chip reliably*, and, thus, *iii)* improve overall system performance while ensuring reliable system operation.

To this end, we first experimentally demonstrate and analyze designed-induced variation in modern DRAM devices by testing and characterizing 96 DIMMs (768 DRAM chips). Our experimental study shows that *i)* modern DRAM chips exhibit design-induced latency variation in both row and column directions, *ii)* access latency gradually increases in the row direction within a DRAM cell array (mat) and this pattern repeats in every mat, and *iii)* some columns require higher latency than others due to the internal hierarchical organization of the DRAM chip.

Our characterization identifies DRAM regions that are *vulnerable* to errors, if operated at lower latency, and finds consistency in their locations across a given DRAM chip generation, due to design-induced variation. Variations in the vertical and horizontal dimensions, together, divide the cell array into heterogeneous-latency regions, where cells in some regions require longer access latencies for reliable operation. Reducing the latency *uniformly across all regions* in DRAM would improve performance, but can introduce failures in the *inherently slower* regions that require longer access latencies for correct operation. We refer to these inherently slower regions of DRAM as design-induced *vulnerable regions*.

Based on our extensive experimental analysis, we develop two mechanisms that reliably reduce DRAM latency. First, DIVA Profiling uses runtime profiling to *dynamically* identify the lowest DRAM latency that does not introduce failures. DIVA Profiling exploits design-induced variation and periodically profiles *only* the *vulnerable regions* to determine the lowest DRAM latency at low cost. It is the first mechanism to *dynamically* determine the lowest latency that can be used to operate DRAM *reliably*. DIVA Profiling reduces the latency of read/write requests by 35.1%/57.8%, respectively, at 55°C. Our second mechanism, DIVA Shuffling, shuffles data such that values stored in vulnerable regions are mapped to multiple error-correcting code (ECC) codewords. As a result, DIVA Shuffling can correct 26% more multi-bit errors than conventional ECC. Combined together, our two mechanisms reduce read/write latency by 40.0%/60.5%, which translates to an overall system performance improvement of 14.7%/13.7%/13.8% (in 2-/4-/8-core systems) over a variety of workloads, while ensuring reliable operation.

## CCS CONCEPTS

• **Computer systems organization** → **Architectures**; **Processors and memory architectures**; **Reliability**; • **Hardware** → **Dynamic memory**;

## KEYWORDS

Memory Systems; DRAM; Latency Variation; Fault Tolerance

**ACM Reference format:**
Donghyuk Lee, Samira Khan, Lavanya Subramanian, Saugata Ghose, Rachata Ausavarungnirun, Gennady Pekhimenko, Vivek Seshadri, and Onur Mutlu. 2017. Design-Induced Latency Variation in Modern DRAM Chips: Characterization, Analysis, and Latency Reduction Mechanisms. In *Proceedings of SIGMETRICS '17, June 5–9, 2017, Urbana-Champaign, IL, USA,* 1 page.
DOI: http://dx.doi.org/10.1145/3078505.3078533

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


*SIGMETRICS '17, June 5–9, 2017, Urbana-Champaign, IL, USA*
 ISBN 978-1-4503-5032-7/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078505.3078533