Demystifying Complex Workload–DRAM Interactions: An Experimental Study

Saugata Ghose[†] Tianshi Li[†] Nastaran Hajinazar[‡†]
Damla Senol Cali[†] Onur Mutlu[§†]

[†]Carnegie Mellon University [‡]Simon Fraser University [§]ETH Zürich

ABSTRACT

It has become increasingly difficult to understand the complex interaction between modern applications and main memory, composed of Dynamic Random Access Memory (DRAM) chips. Manufacturers and researchers are developing many different types of DRAM, with each DRAM type catering to different needs (e.g., high throughput, low power, high memory density). At the same time, the memory access patterns of prevalent and emerging applications are rapidly diverging, as these applications manipulate larger data sets in very different ways. As a result, the combined DRAM–workload behavior is often difficult to *intuitively* determine today, which can hinder memory optimizations in both hardware and software.

In this work, we identify important families of workloads, as well as prevalent types of DRAM chips, and rigorously analyze the combined DRAM–workload behavior. To this end, we perform a comprehensive experimental study of the interaction between nine different DRAM types (DDR3/4, LPDDR3/4, GDDR5, Wide I/O, Wide I/O 2, HBM, HMC) and 115 modern applications and multiprogrammed workloads from six diverse application families (desktop/scientific, server/cloud, multimedia acceleration, network acceleration, GPGPU, OS routines). We draw 12 key observations from our characterization, enabled in part by our development of new metrics that quantify the effect of memory access patterns on hardware utilization. We highlight our five most significant observations here:

(1) Despite having 50% higher memory bandwidth than DDR3, the *newer* DDR4 *rarely outperforms* DDR3 on the applications we evaluate, as DDR4's access latency is 11–14% higher.
(2) The high-bandwidth HMC does *not* outperform DDR3 for most single-thread workloads and many multithreaded applications. This is because HMC's design trade-offs (e.g., a row width that is 97% smaller than DDR3) fundamentally limit opportunities for exploiting spatial locality. For example, single-thread desktop and scientific applications actually perform 5.8% *worse* with HMC than with DDR3, on average, even though HMC offers 87.4% more memory bandwidth. HMC provides significant performance improvements over other DRAM types in cases where application spatial locality is low (or is destroyed), such as highly-memory-intensive multiprogrammed workloads.
(3) While low-power DRAM types typically perform worse than standard-power DRAM for most memory-intensive applications, some low-power DRAM types perform well when bandwidth demand is very high. For example, on average, LPDDR4 performs only 7.0% worse than DDR3 for our multiprogrammed desktop workloads, while consuming 68.2% less energy, and Wide I/O 2 performs 2.3% *better* than DDR3 for multimedia acceleration.
(4) The best DRAM for a heterogeneous system depends heavily on the predominant function(s) performed by the system. We study three types of applications for heterogeneous systems. First, multimedia acceleration benefits most from high-throughput memories that exploit a high amount of *spatial locality*, running up to 21.6% faster with GDDR5 and 14.7% faster with HBM than DDR3, but only 5.0% faster with HMC. Second, a network accelerator's memory requests are highly bursty and do *not* exhibit significant spatial locality, and are thus a good fit for the high bank-level parallelism of HMC (88.4% faster on average over DDR3). Third, GPGPU applications exhibit a wide range of memory intensity, but memory-intensive GPGPU applications typically also take advantage of spatial locality due to memory coalescing, and perform more effectively with HBM (26.9% higher on average over DDR3) and GDDR5 (39.7%) than with DDR3 or HMC.
(5) Several common OS routines (e.g., file I/O, process forking) exhibit extremely high spatial locality, and do not benefit from high amounts of bank-level parallelism. As a result, they perform better with memories such as DDR3 and GDDR5, which have lower access latencies than the other memory types that we study. Since OS routines are used across most computer systems in a widespread manner, we believe DRAM designers must provide low-latency access, instead of the current trend increasing the latency in order to deliver greater throughput.

For more information on our extensive experimental characterization, we refer the reader to the full version of our paper [1]. We hope that the trends we identify can drive optimizations in both hardware and software design. To aid further study, we open-source our extensively-modified simulators [2, 4], as well as MemBen, a benchmark suite containing our applications [3].

KEYWORDS

DRAM; memory systems; experimental characterization; performance modeling; power modeling; workload analysis

ACM Reference Format:
S. Ghose et al. 2019. Demystifying Complex Workload–DRAM Interactions: An Experimental Study. In *ACM SIGMETRICS / International on Measurement & Modeling of Computer Systems (SIGMETRICS '19 Abstracts), June 24–28, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3309697.3331482

REFERENCES

[1] S. Ghose, T. Li, N. Hajinazar, D. Senol Cali, and O. Mutlu, "Demystifying Complex Workload–DRAM Interactions: An Experimental Study," *POMACS*, 2019, available as arXiv:1902.07609 [cs.AR].
[2] SAFARI Research Group, "GPGPUSim+Ramulator — GitHub Repository," https://github.com/CMU-SAFARI/GPGPUSim-Ramulator.
[3] SAFARI Research Group, "MemBen: A Memory Benchmark Suite for Ramulator — GitHub Repository," https://github.com/CMU-SAFARI/MemBen.
[4] SAFARI Research Group, "Ramulator: A DRAM Simulator — GitHub Repository," https://github.com/CMU-SAFARI/ramulator.

SIGMETRICS '19 Abstracts, June 24–28, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6678-6/19/06.
https://doi.org/10.1145/3309697.3331482