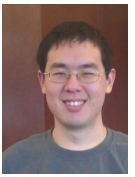
Subarray-Level Parallelism (SALP) in DRAM




Yoongu Kim **Vivek Seshadri** **Donghyuk Lee** **Jamie Liu** **Onur Mutlu**

Modern DRAMs have multiple banks to serve multiple memory requests in parallel. However, when two requests go to the same bank, they have to be served serially, exacerbating the high latency of off-chip memory. Adding more banks to the system to mitigate this problem incurs high system cost. Our goal in this work is to achieve the benefits of increasing the number of banks with a low cost approach. To this end, we propose three new mechanisms that overlap the latencies of different requests that go to the same bank. The key observation exploited by our mechanisms is that a modern DRAM bank is implemented as a collection of subarrays that operate largely independently while sharing few global peripheral structures.

Our proposed mechanisms (SALP-1, SALP-2, and MASA) mitigate the negative impact of bank serialization by overlapping different components of the bank access latencies of multiple requests that go to different subarrays within the same bank. SALP-1 requires no changes to the existing DRAM structure and only needs reinterpretation of some DRAM timing parameters. SALP-2 and MASA require only modest changes (< 0.15% area overhead) to the DRAM peripheral structures, which are much less design constrained than the DRAM core. Evaluations show that all our schemes significantly improve performance for both single-core systems and multi-core systems. Our schemes also interact positively with application-aware memory request scheduling in multi-core systems.

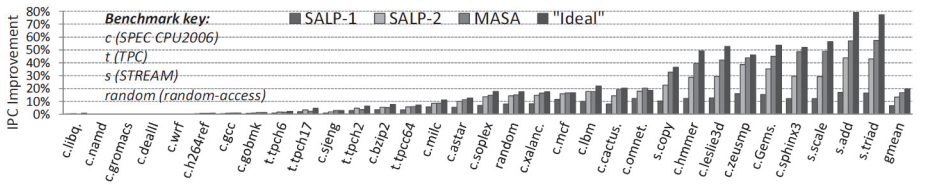


Fig. 1: IPC Improvement (Line-interleaved; Closed-row)

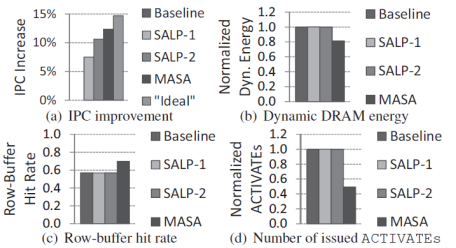


Fig. 2: IPC, Energy, Row-Buffer Hit-Rate, ACTIVATEs (Row-interleaved; Open-row)