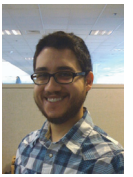
A Case for Small Row Buffers in Non-Volatile Main Memories

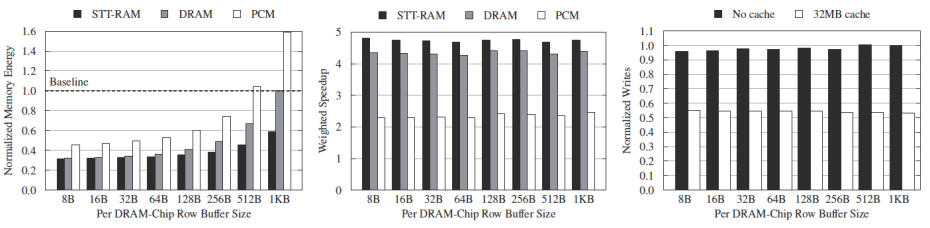



Justin Meza **Onur Mutlu**

DRAM-based main memories have read operations that destroy the read data, and as a result, must buffer large amounts of data on each array access to keep chip costs low. Unfortunately, system-level trends such as increased memory contention in multi-core architectures and data mapping schemes that improve memory parallelism lead to only a small amount of the buffered data to be accessed. This makes buffering large amounts of data on every memory array access energy-inefficient; yet organizing DRAM chips to buffer small amounts of data is costly, as others have shown.

Emerging non-volatile memories (NVMs) such as PCM, STT-RAM, and RRAM, however, do not have destructive read operations, opening up opportunities for employing small row buffers without incurring additional area penalty and/or design complexity. In this work, we discuss and evaluate architectural changes to enable small row buffers at a low cost in NVMs.

We find that on a multi-core system, reducing the row buffer size can greatly reduce main memory dynamic energy compared to a DRAM baseline with large row sizes (Fig. 1a), though there are diminishing marginal returns as row buffer size decreases due to energy consumption being dominated by the energy required to transfer data. Interestingly, for some NVM technologies, such as STT-RAM, performance can be gained by using smaller row buffers which enable a more efficient access protocol that eliminates the precharge delay normally incurred on row buffer misses, and relaxes the tRRD and tFAW timing parameters to enable more banks to be accessed simultaneously (Fig. 1b). We also evaluate our various row buffer size configurations with a 32MB e-DRAM cache, and find that with or without a cache, decreasing the row buffer size has only a small effect on the number of NVM writes performed due to the low row buffer locality present in our multi-core system. In contrast, the addition of a reasonably-sized e-DRAM cache has a large impact on the reduction of writes, decreasing the number of writes by 37% to 47% across the various row buffer sizes (Fig. 1c).



(a) Memory energy with block interleaving. (b) Performance with block interleaving. (c) Writes with and without a 32MB cache.

Fig. 1: Multi-core main memory energy, performance, and writes.