In most modern systems, the memory subsystem is managed and accessed at multiple different granularities (e.g., words, cache lines, and pages) at various resources. In this thesis, we observe that such multi-granularity management results in significant inefficiency in the memory subsystem. Specifically, we observe that 1) page-granularity virtual memory unnecessarily triggers large memory operations, and 2) existing cache-line granularity off-chip memory interface is inefficient for performing bulk data operations and operations that exhibit poor spatial locality. To address these problems, we present a series of techniques in this thesis.

First, to address the inefficiency of existing page-granularity virtual memory systems, we propose a new framework called page overlays. At a high level, our framework augments the existing virtual memory framework with the ability to track a new version of a subset of cache lines within each virtual page. We show that this simple extension is powerful by demonstrating its benefits on a number of different applications.

Second, we show that DRAM, the technology used to build main memory, can be used to perform more complex operations than just store data. We propose RowClone, a mechanism to perform bulk data copy and initialization operations completely inside DRAM, and Buddy RAM, a mechanism to perform bulk bitwise logical operations using DRAM technology. Both these techniques achieve an order-of-magnitude improvement in performance and energy-efficiency of the respective operations.

Third, to improve the performance of non-unit strided access patterns, we propose Gather-Scatter DRAM (GS-DRAM), a technique that exploits the organization of DRAM modules to effectively gather or scatter values with a power-of-2 strided access patterns. For these access patterns, GS-DRAM achieves near-ideal bandwidth and cache utilization, without increasing the latency of fetching data from memory.

Finally, to improve the performance of the protocol to maintain the coherence of dirty cache blocks, we propose the Dirty-Block Index (DBI), a new way of tracking dirty blocks in the on-chip caches. In addition to improving the efficiency of bulk data coherence, DBI has several applications including high-performance memory scheduling, efficient cache lookup bypassing, and enabling heterogeneous ECC for on-chip caches.