Abstract

Memory system design is increasingly influencing modern multi-core architectures from both performance and power perspectives. Both main memory latency and bandwidth have improved at a rate that is slower than the increase in processor core count and speed. Off-chip memory, primarily built from DRAM, has received significant attention in terms of architecture and design for higher performance. These performance improvement techniques include sophisticated memory access scheduling, use of multiple memory controllers, mitigating the impact of DRAM refresh cycles, and so on. At the same time, new non-volatile memory technologies have become increasingly viable in terms of performance and energy. These alternative technologies offer different performance characteristics as compared to traditional DRAM.

With the advent of 3D stacking, on-chip memory in the form of 3D stacked DRAM has opened up avenues for addressing the bandwidth and latency limitations of off-chip memory. Stacked DRAM is expected to offer abundant capacity — 100s of MBs to a few GBs — at higher bandwidth and lower latency. Researchers have proposed to use this capacity as an extension to main memory, or as a large last-level DRAM cache. When leveraged as a cache, stacked DRAM provides opportunities and challenges for improving cache hit rate, access latency, and off-chip bandwidth.

Thus, designing off-chip and on-chip memory systems for multi-core architectures is complex, compounded by the myriad architectural, design and technological choices, combined with the characteristics of application workloads. Applications have inherent spatial locality and access parallelism that influence the memory system response in terms of latency and bandwidth.

In this thesis, we construct an analytical model of the off-chip main memory system to comprehend this diverse space and to study the impact of memory system parameters and workload characteristics from latency and bandwidth perspectives. Our model, called ANATOMY*, uses a queuing network formulation of the memory system parameterized with workload characteristics to obtain a closed form solution for the average miss penalty experienced by the last-level cache. We validate the model across a wide variety of memory configurations on four-core, eight-core and sixteen-core architectures.* ANATOMY *is able to predict memory latency with average errors of* 8.1%*,* 4.1% *and* 9.7% *over quad-core, eight-core and sixteen-core configurations respectively. Further,* ANATOMY *identifies better performing design points accurately thereby allowing architects and designers to explore the more promising design points in greater detail. We demonstrate the extensibility and applicability of our model by exploring a variety of memory design choices such as the impact of clock speed, benefit of multiple memory controllers, the role of banks and channel width, and so on. We also demonstrate* ANATOMY*'s ability to capture architectural elements such as memory scheduling mechanisms and impact of DRAM refresh cycles. In all of these studies,* ANATOMY *provides insight into sources of memory performance bottlenecks and is able to quantitatively predict the benefit of redressing them.*

An insight from the model suggests that the provisioning of multiple small row-buffers in each DRAM bank achieves better performance than the traditional one (large) row-buffer per bank design. Multiple row-buffers also enable newer performance improvement opportunities such as intra-bank parallelism between data transfers and row activations, and smart row-buffer allocation schemes based on workload demand. Our evaluation (both using the analytical model and detailed cycle-accurate simulation) shows that the proposed DRAM reorganization achieves significant speed-up as well as energy reduction.

Next we examine the role of on-chip stacked DRAM caches at improving performance by reducing the load on off-chip main memory. We extend ANATOMY *to cover DRAM caches.* ANATOMY-Cache *takes into account all the key parameters/design issues governing DRAM cache organization namely, where the cache metadata is stored and accessed, the role of cache block size and set associativity and the impact of block size on row-buffer hit rate and off-chip*

bandwidth. Yet the model is kept simple and provides a closed form solution for the average miss penalty experienced by the last-level SRAM cache. ANATOMY-Cache *is validated against detailed architecture simulations and shown to have latency estimation errors of* 10.7% *and* 8.8% *on average in quad-core and eight-core configurations respectively. An interesting insight from the model suggests that under high load, it is better to bypass the congested DRAM cache and leverage the available idle main memory bandwidth. We use this insight to propose a refresh reduction mechanism that virtually eliminates refresh overhead in DRAM caches. We implement a low-overhead hardware mechanism to record accesses to recent DRAM cache pages and refresh only these pages. Older cache pages are considered invalid and serviced from the (idle) main memory. This technique achieves average refresh reduction of* 90% *with resulting memory energy savings of* 9% *and overall performance improvement of* 3.7%*.*

Finally, we propose a new DRAM cache organization that achieves higher cache hit rate, lower latency and lower off-chip bandwidth demand. Called the Bi-Modal Cache*, our cache organization brings three independent improvements together: (i) it enables parallel tag and data accesses, (ii) it eliminates a large fraction of tag accesses entirely by use of a novel* way locator *and (iii) it improves cache space utilization by organizing the cache sets as a combination of some big blocks (512B) and some small blocks (64B). The* Bi-Modal Cache *reduces hit latency by use of the way locator and parallel tag and data accesses. It improves hit rate by leveraging the cache capacity efficiently – blocks with low spatial reuse are allocated in the cache at 64B granularity thereby reducing both wasted off-chip bandwidth as well as cache internal fragmentation. Increased cache hit rate leads to reduction in off-chip bandwidth demand. Through detailed simulations, we demonstrate that the Bi-Modal Cache achieves overall performance improvement of* 10.8%*,* 13.8% *and* 14.0% *in quad-core, eight-core and sixteen-core workloads respectively over an aggressive baseline.*