Staged Memory Scheduling: Achieving high Performance and Scalability in Heterogeneous Systems


Rachata Ausavarungnirun


Kevin Chang


Lavanya Subramanian


Gabriel H. Loh


Onur Mutlu

When multiple processor (CPU) cores and a GPU integrated together on the same chip share the off-chip main memory, requests from the GPU can heavily interfere with requests from the CPU cores, leading to low system performance and starvation of CPU cores. Unfortunately, state-of-the-art application-aware memory scheduling algorithms are ineffective at solving this problem at low complexity due to the large amount of GPU traffic. A large and costly request buffer is needed to provide these algorithms with enough visibility across the global request stream, requiring relatively complex hardware implementations.

This paper proposes a fundamentally new approach that decouples the memory controller's three primary tasks into three significantly simpler structures that together improve system performance and fairness, especially in integrated CPU-GPU systems. Our three-stage memory controller first groups requests based on row-buffer locality. This grouping allows the second stage to focus only on inter-application request scheduling. These two stages enforce high-level policies regarding performance and fairness, and therefore the last stage consists of simple per-bank FIFO queues (no further command reordering within each bank) and straightforward logic that deals only with low-level DRAM commands and timing.

We evaluate the design trade-offs involved in our Staged Memory Scheduler (SMS) and compare it against three state-of-the-art memory controller designs. Our evaluations show that SMS improves CPU performance without degrading GPU frame rate beyond a generally acceptable level, while being significantly less complex to implement than previous application-aware schedulers. Furthermore, SMS can be configured by the system software to prioritize the CPU or the GPU at varying levels to address different performance needs.

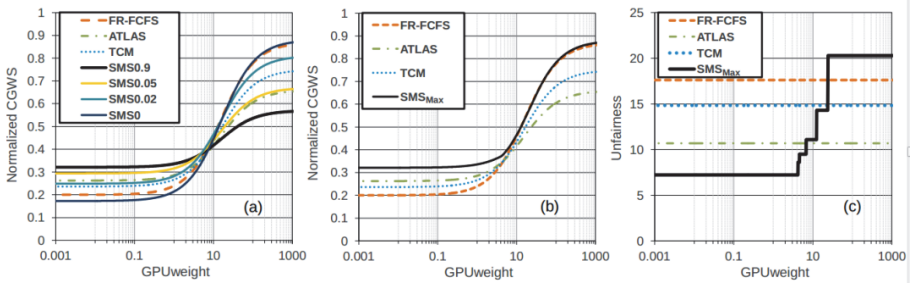

Figure 1: Combined performance of the CPU-GPU system (a) for different SMS configurations with different SJF probability, (b) for the SMS configuration sthat provides the maximum CGWS, and (c) unfairness for the configuration that provides the maximum CGWS.