Observations and Opportunities in Architecting Shared Virtual Memory for Heterogeneous Systems

Jan Vesely
Dept. of Computer Science

11/9/2016 at 10:15 am
CoRE B (305)

Abstract

Computing is becoming increasingly heterogeneous with accelerators like GPUs being tightly integrated with CPUs on the same die. Extending the CPUs virtual addressing mechanism to these accelerators is a key step in making accelerators easily programmable. In this work, we analyze, using real-system measurements, shared virtual memory across the CPU and an integrated GPU. We make several key observations and highlight consequent research opportunities: (1) servicing a TLB miss from the GPU can be an order of magnitude slower than that from the CPU and consequently it is imperative to enable many concurrent TLB misses to hide this larger latency; (2) divergence in memory accesses impacts the GPUs address translation more than the rest of the memory hierarchy, and research in designing address translation mechanisms tolerant to this effect is imperative; and (3) page faults from the GPU are considerably slower than that from the CPU and software-hardware co-design is essential for efficient implementation of page faults from throughput-oriented accelerators like GPUs. We present a detailed measurement study of a commercially available integrated APU that illustrates these effects and motivates future research opportunities.