Memory Coalescing for Hybrid Memory Cube

Xi Wang
Texas Tech University
xi.wang@ttu.edu

John D. Leidel
Tactical Computing Laboratories
jleidel@tactcomplabs.com

Yong Chen
Texas Tech University
yong.chen@ttu.edu

ABSTRACT

Arguably, many data-intensive applications pose significant challenges to conventional architectures and memory systems, especially when applications exhibit non-contiguous, irregular, and small memory access patterns. The performance of data-intensive workloads running on traditional processors and architectures is not as expected, especially when encountering the irregular memory-access patterns that produce random memory footprints. The long memory access latency can dramatically slow down the overall performance of applications. In addition, the growing data-level parallelism in these data-intensive workloads increase the concurrency in data accesses and further complicate the memory system support.

In order to reduce the latency of memory accesses, two main research directions exist in current studies. The first direction focuses on moving computation to data, thus reducing the need and the memory traffic of moving data between memory and processors. These efforts include near data processing (NDP) and processing in memory (PIM). While these techniques bridge the gap between the processors and main memory, they cannot eliminate memory accesses, thus the bandwidth of conventional DDR devices restricts the ceiling of the overall performance.

The second direction focuses on developing advanced memory devices that satisfy the growing desire of high memory bandwidth and low latency access. This motivation stimulates the advent of novel 3D-staked memory devices such as the Hybrid Memory Cube (HMC), which provides significantly higher bandwidth compared with the conventional JEDEC DDR devices.

Even though many existing studies have been devoted to achieving high bandwidth throughput of HMC, the bandwidth potential cannot be fully exploited due to the lack of highly efficient interfacing methodology designed and optimized for HMC devices. As the size of the memory request transactions is consistent with the cache line size and the fixed burst size in DDR interface, the memory request sizes are usually fixed as 64B in mainstream architectures. It may evoke higher latency and control overhead to apply the same policy to the flexible packet-based HMC devices naively.

In this research, we introduce a novel memory coalescer methodology that facilitates memory bandwidth efficiency and the overall performance through an efficient and scalable memory request coalescing interface for HMC. We present the design and implementation of this approach on RISC-V embedded cores with attached HMC devices. Our evaluation results show that the new memory coalescer eliminates 47.47% memory accesses to HMC and improves the overall performance by 13.14% on average.

ICPP 2018 August 13 – 16, Eugene, Oregon
 ISBN 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4