Exploring Memory Coalescing for 3D-Stacked Hybrid Memory Cube

Xi Wang, John D. Leidel, Yong Chen

Department of Computer Science, Texas Tech University

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 long memory access latency can dramatically slow down the overall performance of applications. The growing desire of high memory bandwidth and low latency access stimulate 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. 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.

Workflow

1. Pipelined Request Sorting Network

- Fully pipelined HMC controller overlaps the latency of memory accesses
- We construct a pipelined request sorting network to hide the coalescing latency.

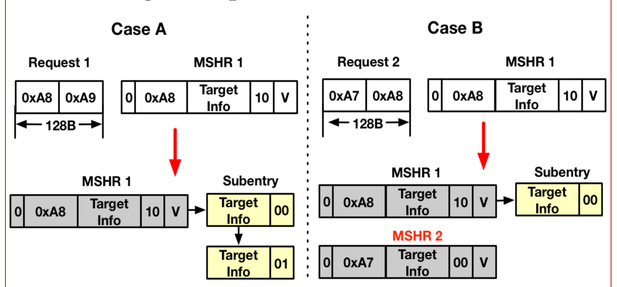
2. DMC Unit

- A two-phase coalescing model is introduced in this work.
- DMC unit is responsible for the first-phase memory coalescing, which constructs large request packets.
- Combined requests are immediately pushed into the coalesced request queue (CRQ),

3. Dynamic MSHRs

- Dynamic MSHRs are responsible for the second phase coalescing through the request merging in the MSHR entries.
- Requested cache line is derived by: ***Subentry.addr = Entry.addr + Line.ID × Line.size***

Coalescing Example



Introduction

In the epoch of big data, many data-intensive applications with intense memory traffic such as data mining, machine learning, pattern recognitions, video processing, etc., pose significant challenges to memory systems.

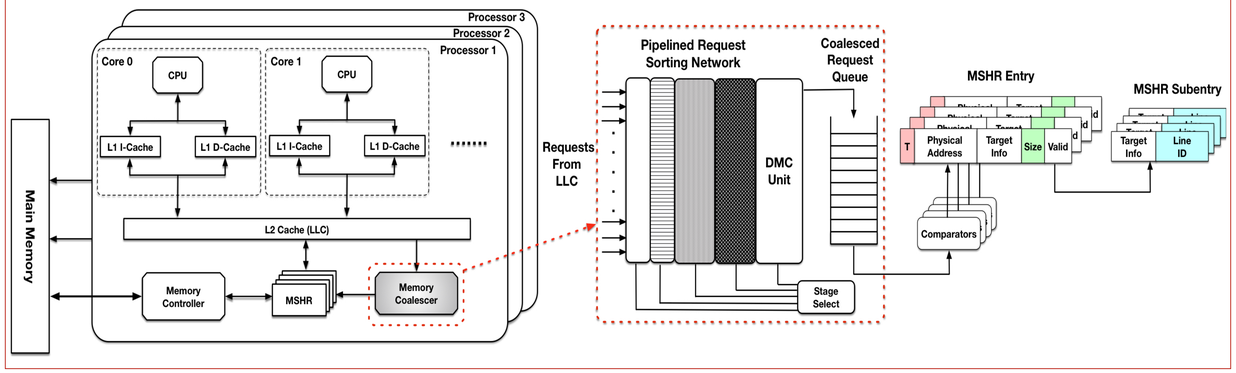
- ***Growing data-level parallelism triggers more frequent data accesses***.
- ***Irregular memory-access patterns causes much higher cache miss rate***.

A new 3D-stacked memory device named Hybrid Memory Cube (HMC) is designed to satisfy the desire of growing bandwidth [1]. While, the maximum throughput (320 GB/s) of HMC is achieved through the transactions of large and flexible request packets [2, 3].

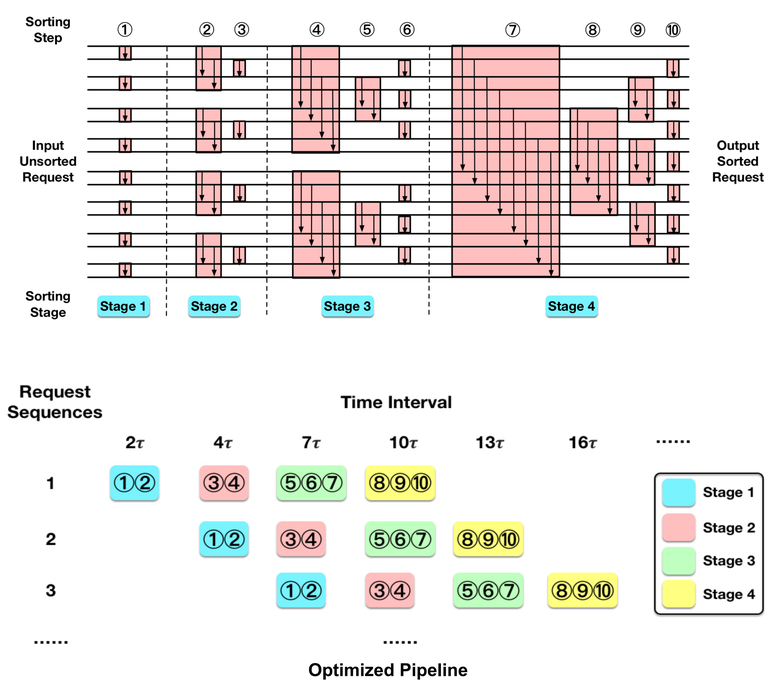
- DDR interface forces fixed request size based on the size of the cache line.

Therefore, a novel memory coalesing interface for HMC is introduced in this work to address the limited applicability of existing memory interface and alleviate the performance penalty.

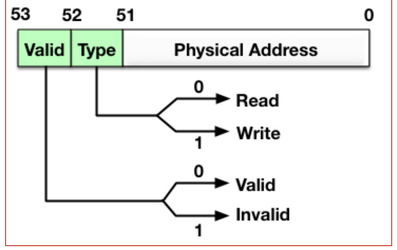
Architecture



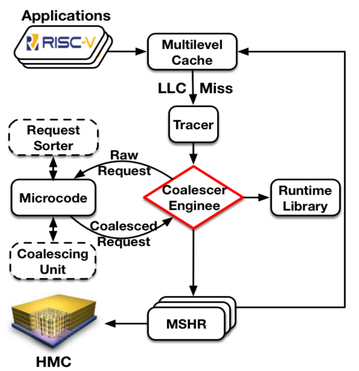
Pipelined Request Sorting Network



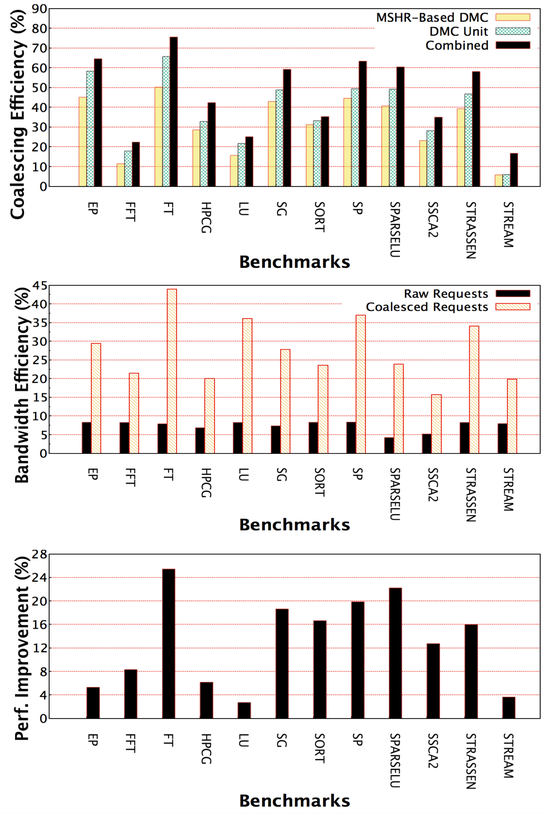
Address Extensions



Implementations



Evaluations



Conclusion

- In this study, we have presented a novel memory coalescer infrastructure for HMC.
- The memory coalescer largely reduces the memory access latency and boosts the bandwidth efficiency.
- The evaluation shows that the memory coalescer eliminated an average of 47.47% memory accesses
- The overall performance is improved by 13.14% on average.

References:

[1] HMC Specification 2.1. Technical report, December 2015.

[2] Xi Wang, John D. Leidel, Yong Chen, Memory Coalescing for Hybrid Memory Cube. In *ICPP 2018, Eugene, Oregon, USA.*

[3] Xi Wang, John D. Leidel, Yong Chen, Concurrent Dynamic Memory Coalescing on GoblinCore-64 Architecture. In *MEMSYS 2016, Washington, DC, USA.*