Concurrent Dynamic Memory Coalescing for Data-Intensive Computing

Xi Wang, PhD Student
Advisor: Dr. Yong Chen
Computer Science Department
Texas Tech University

Abstract

The majority of modern microprocessors are architected to utilize multi-level data caches as a primary optimization to reduce the latency and increase the perceived bandwidth from an application. The spatial and temporal locality provided by data caches work well for applications that access memory in a linear or regular fashion. However, many growingly critical data-intensive applications often exhibit random or irregular, non-deterministic memory access patterns, which induce a significant number of data cache misses, and reduce the natural performance benefit from the data cache. In response to the performance penalties inherently present with many data-intensive applications, we have constructed a unique memory hierarchy within the GoblinCore-64 (GC64) architecture that is explicitly designed as a scalable, open architecture for data-intensive computing. The GC64 architecture combines a RISC-V-based core coupled with latency-hiding architectural features to a memory hierarchy with Hybrid Memory Cube (HMC) devices. In order to cope with the inherent non-determinism of applications and to exploit the packetized interface presented by the HMC device, we develop a methodology and associated implementation of a dynamic memory coalescing (DMC) unit that permits us to statistically sample memory requests from non-deterministic applications and coalesce them into the largest possible HMC payload requests. We also propose a tree-based concurrent DMC model that groups the memory accesses by their memory partitions. This DMC model is designed, implemented, and evaluated in the MMU of RISC-V ISA targeting at HMC. In this talk, I will introduce the detailed algorithms and design of our DMC model. I will present our evaluations that validate the efficacy of our design, as well as comparisons between our methodology and existing work.

Bio

Xi Wang is a Ph.D. student of Dr. Yong Chen in the Computer Science Department at Texas Tech University (TTU), US. He received his M.S. degree in Computer Science at TTU. He works in the Data-Intensive Scalable Computing Laboratory (DISCL) and his research area is High Performance Computing (HPC) with a focus on memory systems and computer architecture design. In particular, his current work is about boosting the performance of memory coalescing through advanced dynamic memory coalescing (DMC) techniques in HPC.