World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:18, No:09, 2024

Hydrogen: Contention-Aware Hybrid Memory Management for Heterogeneous CPU-GPU Architectures

Authors : Yiwei Li, Mingyu Gao

Abstract : Integrating hybrid memories with heterogeneous processors could leverage heterogeneity in both compute and memory domains for better system efficiency. To ensure performance isolation, we introduce Hydrogen, a hardware architecture to optimize the allocation of hybrid memory resources to heterogeneous CPU-GPU systems. Hydrogen supports efficient capacity and bandwidth partitioning between CPUs and GPUs in both memory tiers. We propose decoupled memory channel mapping and token-based data migration throttling to enable flexible partitioning. We also support epoch-based online search for optimized configurations and lightweight reconfiguration with reduced data movements. Hydrogen significantly outperforms existing designs by 1.21x on average and up to 1.31x.

Keywords : hybrid memory, heterogeneous systems, dram cache, graphics processing units
Conference Title : ICCD 2024 : International Conference on Computer Design
Conference Location : Paris, France
Conference Dates : September 16-17, 2024