CoNDA: Efficient Cache Coherence Support for Near-Data Accelerators
Amirali Boroumand, Saugata Ghose, Minesh Patel, Hasan Hassan, Brandon Lucia, Rachata Ausavarungnirun, Kevin Hsieh, Nastaran Hajinazar, Krishna Malladi, Hongzhong Zheng, Onur Mutlu

Coherence For NDAs

Challenge: Coherence between NDAs and CPUs
1. Large cost of on-chip communication
2. NDAs generate large off-chip data movement

It is impractical to use traditional coherence protocols

Analysis of Existing Coherence Mechanisms

Non-Cacheable Approach
- Mark the NDA as non-cacheable
  - Generates a large number of off-chip accesses
  - Significantly hurts CPU threads performance
  - NC fails to provide any energy saving and performs 6.0% worse than CPU-only

Coarse-Grained Coherence
- Get coherence permission for the NDA region
  - Unnecessarily flushes a large amount of dirty data
  - CG fails to provide any performance benefit of NDA and performs 0.4% worse than CPU-only

Fine-Grained Coherence
- Using fine-grained coherence has two benefits:
  1. Simplifies NDA programming model
  2. Allows us to get permissions for only the pieces of data that are actually accessed

CoNDA
We propose CoNDA, a mechanism that uses optimistic NDA execution to avoid unnecessary coherence traffic

Identifying Coherence Violations

Architecture Support

High Level Architecture of CoNDA

Optimistic Execution

Per-word dirty bit mask to mark all uncommitted data updates

CoNDA consistently retains most of Ideal-NDA’s benefits, coming within 19.4% of the Ideal-NDA performance

Evaluation

CoNDA significantly reduces energy consumption and comes within 4.4% of Ideal-NDA