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
Specialized accelerators are now everywhere!

Recent advancement in 3D-stacked technology enabled **Near-Data Accelerators (NDA)**
Coherence For NDAs

Challenge: Coherence between NDAs and CPUs

(1) Large cost of off-chip communication

(2) NDA applications generate a large amount of off-chip data movement

It is impractical to use traditional coherence protocols
Existing Coherence Mechanisms

We extensively study existing NDA coherence mechanisms and make three key observations:

1. These mechanisms **eliminate** a significant portion of NDA’s benefits.

2. The **majority of off-chip coherence traffic** generated by these mechanisms is **unnecessary**.

3. Much of the **off-chip traffic** can be **eliminated** if the **coherence mechanism** has **insight** into the **memory accesses**.
An Optimistic Approach

We find that an optimistic approach to coherence can address the challenges related to NDA coherence.

1. Gain insights before any coherence checks happens.

2. Perform only the necessary coherence requests.
We propose **CoNDA**, a mechanism that uses optimistic NDA execution to avoid unnecessary coherence traffic.

Coherence Resolution:
- Offload NDA kernel
- No Coherence Request
- Send signatures
- Commit or Re-execute

**CPU**:
- Signature

**NDA**:
- Signature

**SAFARI**

**CPU**:
- Concurrent CPU + NDA Execution
- CPU Thread Execution

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

CoNDA comes within 10.4% and 4.4% of performance and energy of an ideal NDA coherence mechanism.
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

SAFARI  Carnegie Mellon  SAMSUNG
SFU  ETH Zürich