PIM-Enabled Instructions: A Low-Overhead, Locality-Aware PIM Architecture

Junwhan Ahn, Sungjoo Yoo, Onur Mutlu\textsuperscript{+}, and Kiyoung Choi

Seoul National University \hspace{1cm} \textsuperscript{+}Carnegie Mellon University
Challenges in Processing-in-Memory

**Cost-effectiveness**
- DRAM die
- Complex Logic

**Programming Model**
- Host Processor
  - Thread
  - Thread
  - Thread
  - Thread
  - Thread
  - Thread
  - Thread
  - Thread
  - Thread

**Coherence & VM**
- Host Processor
- DRAM die
PIM-Enabled Instructions

• Our direction: simple PIM operations as ISA extension
  – Simple: low-overhead implementation
  – ISA extension: No changes to existing programming models
  – *One more thing*: locality-aware dynamic PIM execution
    • Adaptation between host-side and memory-side execution

• Evaluation highlight
  – 47% speedup over conventional systems in large inputs
  – 32% speedup over PIM-only systems in small inputs
  – Impact of data locality, energy efficiency, and more…

Session 6A: Memory Systems I (10:20~10:45)