AXI-PACK: Near-Memory Bus Packing for Bandwidth-Efficient Irregular Workloads

Chi Zhang¹, Paul Scheffler¹, Thomas Benz¹, Matteo Perotti¹, Luca Benini¹²
¹Integrated Systems Laboratory, ETH Zurich; ²DEI, University of Bologna

1 Challenge: irregular memory accesses

- Applications
  - Graph analytics
  - Fluid dynamics
  - Recommender systems

- Challenge to processors
  - Poor bus utilization
  - Cache trashing
  - Long latencies

2 State-of-art solutions

- Core-side stream ISA extensions¹
  - Decouple computing and memory access: hide latency
    - Inherent inefficiency of narrow bus accesses
    - High index-fetching overhead and bus traffic

- Memory-side extensions²
  - Prefetch and pack irregular elements at memory-side
    - Not well co-integrated with Core-side
    - Non-standard solutions

3 Proposal: AXI-Pack on-chip protocol

- Extends Advance eXtensible Interface4 (AXI4)³ on-chip protocol
  - Core-side issues
    - Pattern-aware requests
  - Memory-side responses
    - Densely packed stream

- Features
  - End-to-end irregular streaming
  - Process-In-Memory
  - Bus-packing
  - Standard based
  - Backward compatible
  - Transparent
  - Scalable

4 Design work

- Define AXI4 user extension (7% more bits)
- Extend a RISC-V vector processor
- Design an AXI-Pack adapter for banked Memory

5 Results

- Speedup irregular workloads
  - 5.4x (stride)
  - 2.4x (indirect)

- light-weight and scalable
  - 6.2% extension area

- Improve energy efficiency
  - 5.3x (stride)
  - 2.1x (indirect)

References