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

Chi Zhang  
ETH Zurich  
<chizhang@iis.ee.ethz.ch>

Paul Scheffler  
ETH Zurich  
<paulsc@iis.ee.ethz.ch>

Thomas Benz  
ETH Zurich  
<tbenz@iis.ee.ethz.ch>

Matteo Perotti  
ETH Zurich  
<mperotti@iis.ee.ethz.ch>

Luca Benini  
ETH Zurich/University of Bologna  
<lbenini@iis.ee.ethz.ch>
Introduction

• Data-intensive applications with irregular memory access
  • Applications
    • Graph analytics
    • Fluid dynamics
    • Recommender systems
  • Large and sparse datasets
  • Common irregular memory access patterns\(^1\)
    • Strided
    • Indirect (Gather)

• Challenges to processors
  • Poor utilization of bus bandwidth
  • Frequent cache trashing
  • Long latencies

\(^1\) Z. Wang and T. Nowatzki, “Stream-based memory access specialization for general purpose processors,” 2019 ACM/IEEE 46th Annu. Int. Symp. Comput. Architecture (ISCA), pp. 736–749, 2019.
State-of-art Solutions

• Core-side stream ISA extensions
  • Decouple computing and memory access
  • Drawbacks
    • Ignore downstream interconnects and memory systems
    • High index-fetching overhead and bus traffic
    • Inherent inefficiency of narrow bus accesses

• Memory-side extensions
  • Prefetch and pack irregular elements on the memory side
  • Drawbacks
    • Not well co-integrated with Core-side
    • Occupy virtual or physical memory
    • Ahead-of-time invocation
    • Non-standard solutions
Proposal

• **AXI-PACK**: on-chip protocol for high bandwidth irregular access
  • **Extension** of Advance eXtensible Interface4 (AXI4) on-chip protocol
  • In AXI-Pack:
    • Core-side issues **pattern-aware requests**
    • Memory-side response **densely packed data stream**
  • **Connect** core-side and memory-side extensions

• **AXI-PACK Novelties**
  • **End-to-end** irregular memory streaming
  • **Process-In-Memory** protocol
  • **Bus-packing** enables high bus utilization
  • Based on **standard** protocol
  • **Backward compatibility** and **transparency**
  • **Scalable** for multi-master, multi-slave
Design Work & Results

• Design work
  • Define AXI4 user extension for AXI-PACK
    • ~7% extension
  • Extend a RISC-V vector processor
  • Design an AXI-Pack adapter for banked Memory

• Results
  • Speedup irregular workloads
    • 5.4x (stride) and 2.4x (indirect)
  • light-weight and scalable
    • 6.2% area of RSIC-V vector processor
  • Save energy
    • 5.3x (stride) and 2.1x (indirect) energy efficiency