Near-Memory Parallel Indexing and Coalescing: Enabling Highly Efficient Indirect Access for SpMV

Chi Zhang[1], Paul Scheffler[1], Thomas Benz[1], Matteo Perotti[1], Luca Benini[1,2]
[1]Integrated Systems Laboratory, ETH Zurich; [2]DEI, University of Bologna

1 SpMV Challenges on General-Purpose Architecture

- **Indirect addressing**

Complicates access flows

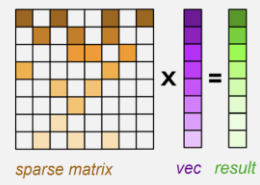
Low computation utilization

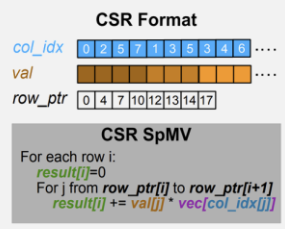
- **Irregular memory access patterns**

Cache trashing & pollution

Low bandwidth efficiency

Long latencies





2 Near-Memory Indirect Stream to the Rescue

- **Handle indirect accesses near memory**

Stream indirect elements from main memory

Smart Memory Controllers[1]

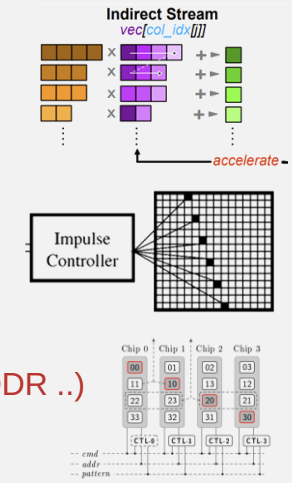
Scatter-Gather DRAM[2]

- **Previous solutions**

Rely on low-granularity (64b) channel

No good solution to modern DRAMs (HBM, LPDDR ..)

~512b access granularity



Can we efficiently stream indirect accesses from modern DRAMs without introducing large caches?

3 Our Proposal: MLP + Coalescence

- **To efficiently handle Indirect stream on modern DRAMs**

Leverage memory-level parallelism (MLP) of indirect stream

Leverage parallel coalescence of narrow and sparse accesses

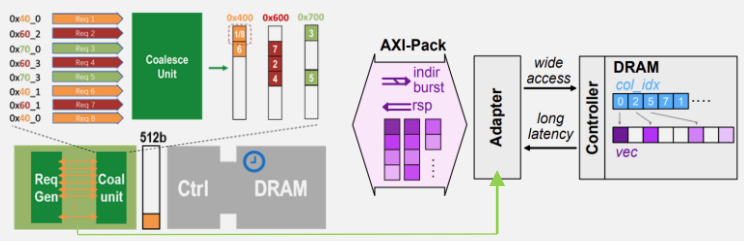
- **Implement AXI-Pack adapter for efficient indirect stream from DRAMs**

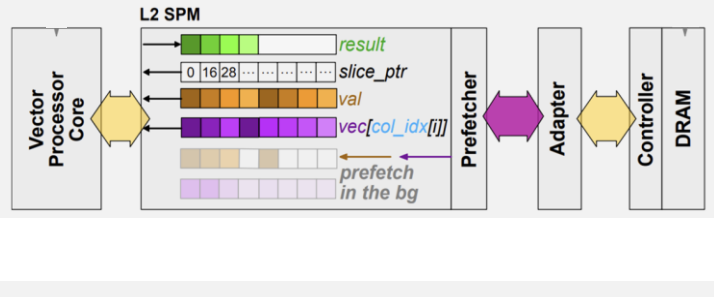
AXI-Pack[3] is a recently proposed extension to AXI4 on-chip protocol

AXI-Pack Support stride and indirect bursts

We extend AXI-Pack adapter's indirect stream unit

Integrate into an open-source RISC-V vector processor system





5 Results

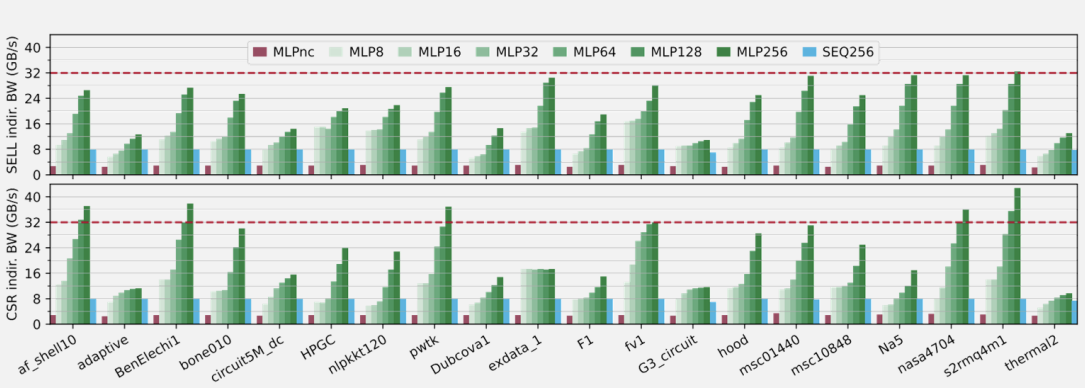
8× increase in **indirect access bandwidth**

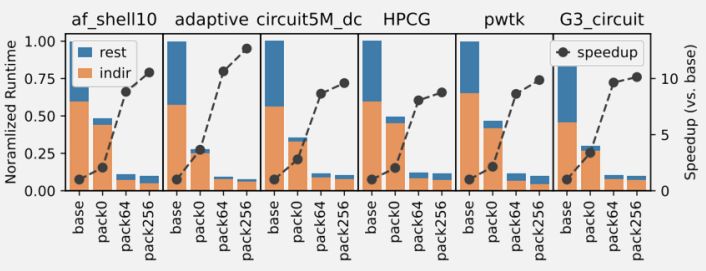
10× speedup in **SpMV performance**

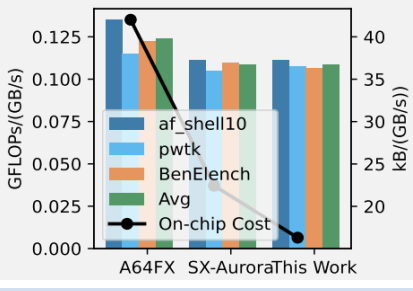
vs. baseline RISCV vector processor with 1MB LLC

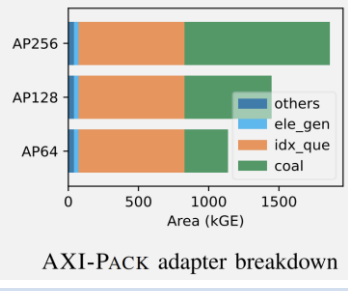
2.6× **superior on-chip efficiency**

vs. SoA vector processors









AXI-Pack adapter breakdown

References

1. Carter, John, et al. "Impulse: Building a smarter memory controller." Proceedings Fifth International Symposium on High-Performance Computer Architecture. IEEE, 1999.
2. Seshadri, Vivek, et al. "Gather-scatter DRAM: In-DRAM address translation to improve the spatial locality of non-unit strided accesses." Proceedings of the 48th International Symposium on Microarchitecture. 2015.
3. Zhang, Chi, et al. "AXI-pack: Near-memory bus packing for bandwidth-efficient irregular workloads." 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2023.