Accelerating Irregular Workloads with Cooperating Indexed Stream Registers

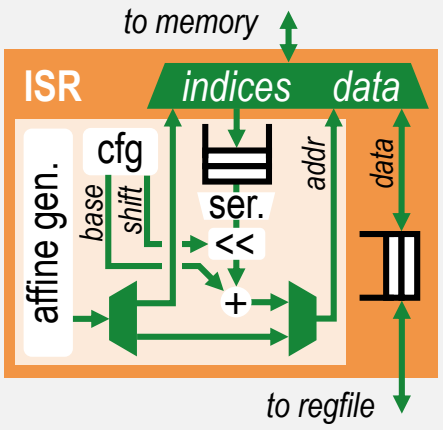
Paul Scheffler[1], Luca Benini[1,2]
[1]Integrated Systems Laboratory, ETH Zurich
[2]Department of Electrical, Electronic, and Information Engineering, University of Bologna

1 Motivation

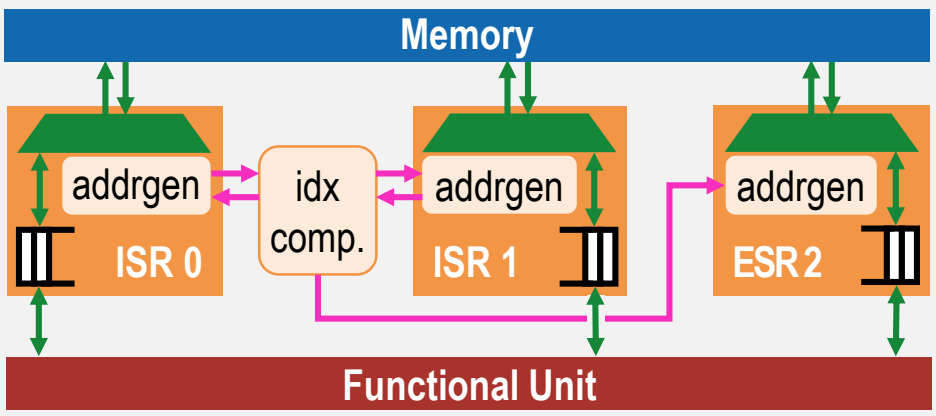
- Irregular workloads (e.g. sparse ML and graph analytics) are **inefficiently handled** in SoA architectures[1]
- CPU, GPU, and accelerator hardware proposals **fall short** in terms of generality or hardware overhead[2]
- We propose and evaluate a RISC-V **stream register (SR) extension accelerating irregular workloads**

2 Architecture

- **Backward-compatible** to existing affine SR design[3]
  - Register accesses *implicitly* push or pop streams
- ***Indexed SR* (ISR)[2]:** Extend SR with **streaming indirection:**
  - Existing affine generator fetches **packed indices** (8 .. 64b) from memory
  - Programmable **index shift** to access higher data axes or structs
  - Unlike RVV indexed ops: indices reside in memory



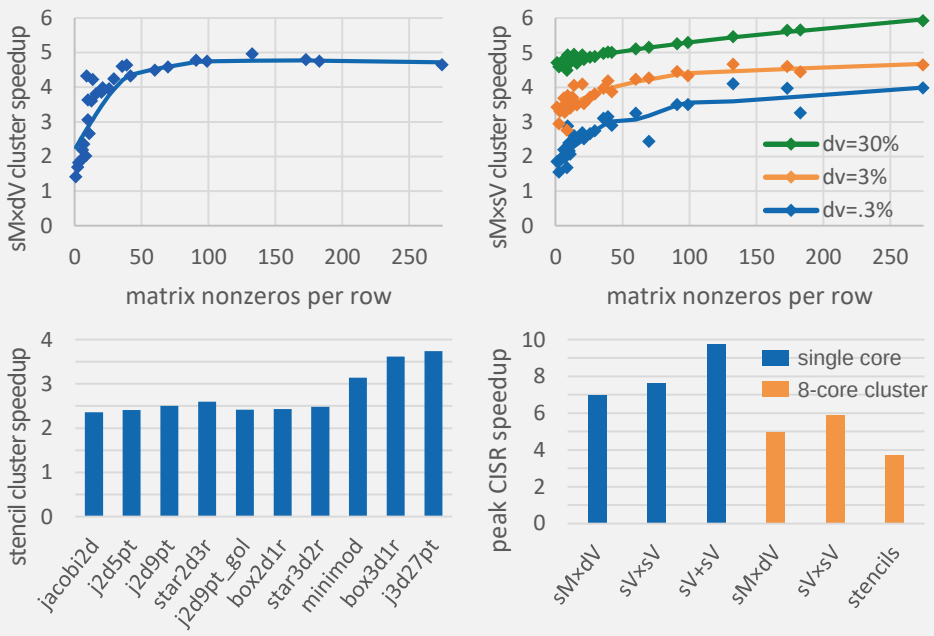
  - ➢ Sparse-dense LA, stencils, compressed data streams…
- ***Cooperating ISRs* (CISRs)[2]***:* 2-3 SRs *exchange indices* for **intersection** or **union** of index streams:



- Assume operands based on *sparse vectors* encoded as a **value + index array pair** (e.g. CSR, CSF)
- **Index comparator** between two ISRs *masks* elements or *injects* zero elements in *trailing* index stream
- Optional **Egress SR (ESR)** writes out sparse array pair
- ➢ Sparse-sparse LA, graph pattern mining…

3 Results

- Evaluation in eight-core RV32 **Snitch cluster**[4]
- GF12LP+ implementation: **+1.8% area** over cluster with affine SRs[3]**, no timing impact**
- Performance evaluation against RV32G in **RTL simulation:**



- *Single core*: up to **7.0×** and **9.8×** speedups on sparse-dense and sparse-sparse LA
- *Cluster:* up to **5.0×**, **5.9×**, **3.7×** speedups on sparse-dense LA, sparse-sparse LA, and stencils
- *Cluster:* up to **93%** FPU utilization, **3.0×** less energy

4 Conclusion

- CISRs extend affine SRs with hardware **indirection**, **intersection**, and **union** to accelerate irregular workloads
- Enable multicore speedups and energy savings of up to **5.9×** and **3.0×** over RV32G and FPU utilizations of up to **93%**
- CISRs could provide a **first step toward formal RISC-V extensions** for SRs and irregular workloads

References

1. *June 2022 HPCG Results*. https://www.hpcg-benchmark.org/custom/index.html%3Flid=155&slid=313.html.
2. P. Scheffler, F. Zaruba, F. Schuiki, T. Hoefler, and L. Benini, “Sparse Stream Semantic Registers: A Lightweight ISA Extension Accelerating General Sparse Linear Algebra”. 2023. Available: https://arxiv.org/abs/2305.05559.
3. F. Schuiki, F. Zaruba, T. Hoefler and L. Benini, “Stream Semantic Registers: A Lightweight RISC-V ISA Extension Achieving Full Compute Utilization in Single-Issue Cores”. In: *IEEE Trans. Comput. 70* (2021), pp. 212–227.
4. F. Zaruba, F. Schuiki, T. Hoefler and L. Benini, et al. "Snitch: A Tiny Pseudo Dual-Issue Processor for Area and Energy Efficient Execution of Floating-Point Intensive Workloads". In: *IEEE Trans. Comput. 70* (2021), pp. 1845–1860.

pulp-platform.org pulp-platform @pulp_platform paulsc@iis.ee.ethz.ch

RISC-V Summit Europe 2023