Accelerating Irregular Workloads with Cooperating Indexed Stream Registers

1 Motivation

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

2 Architecture

- **Backward-compatible** to existing affine SR design.
- **Indexed SR (ISR)**: 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)**: 2-3 SRs exchange indices for intersection or union of index streams.

3 Results

- Evaluation in eight-core RV32 Snitch cluster.
  - GF12LP+ implementation: +1.8% area over cluster with affine SRs, **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 **intersection**, **union**, and **indirection** 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