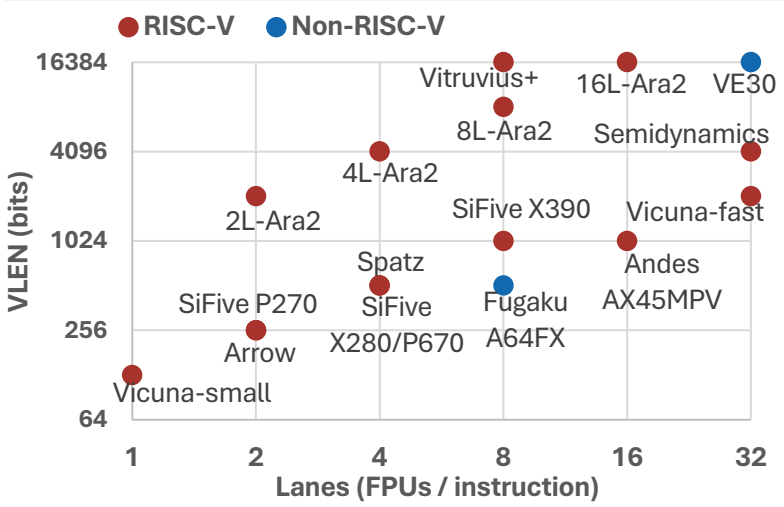
AraXL: A Physically Scalable, Ultra-Wide RISC-V Vector Processor Design for Fast and Efficient Computation on Long Vectors

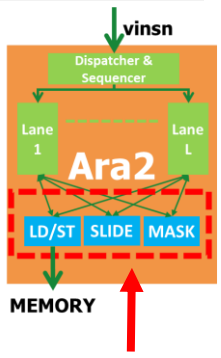
Navaneeth Kunhi Purayil†, Matteo Perotti†, Tim Fischer†, Luca Benini†‡
†ETH Zurich, ‡University of Bologna

1) Scalability Challenge of Vector Processors

- Scaling to more lanes and larger Vector Register Files a challenge.
- SoA processors limited to **16 Kibits and 32 lanes**

RISC-V Vector ISA allows up to 64 Kibits of register length

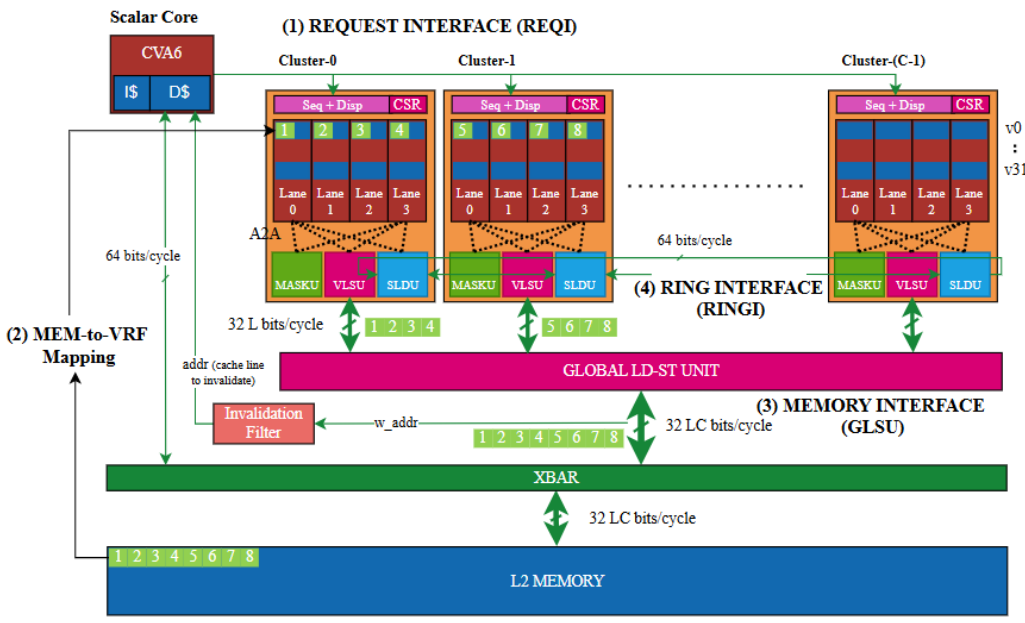




L² complexity (All-2-All interconnect)

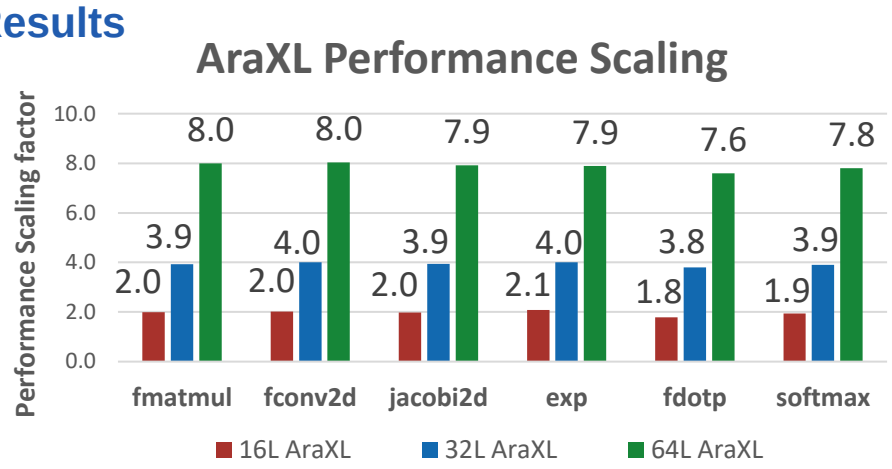
Utilize latency tolerance of long vectors[1] ⟹ Scalable architecture

2) AraXL Architecture



- **Parametric 4-lane Ara2² instances**
- **Low-cost pipelined scalable interconnects**
  (1) REQI (2) RINGI (3) GLSU

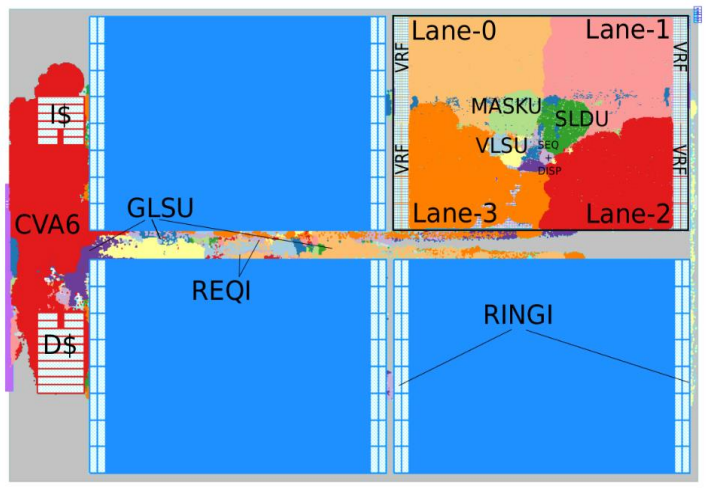
3) Results



- **Linear Performance Scaling** up to 64 lanes
- **High Latency tolerance** < 2%
  at long vector lengths for all interfaces

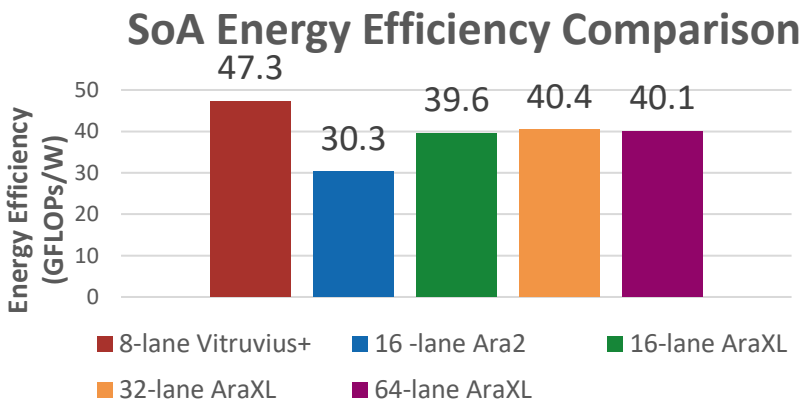
8 → 64 lanes ⟹ 8x Performance

4) Physical Implementation



Hierarchical implementation in GF22 nm FD-SOI technology.

High Energy Efficiency 40 GFLOPs/W @ 1.15 GHz, TT/0.8V/25C



5) Achievements

- ✓ **Novel RISC-V vector processor architecture achieving 64 FPUs (2x SoA)**
- ✓ **Achieves largest vector register length VLEN of 64 Kibits (4x SoA)**
- ✓ **Achieves 146 GFLOPS, 40 GFLOPs/W (Matrix Multiplication >99% FPU utilization)**

References

1. P. Vizcaino et al., “Short Reasons for Long Vectors in HPC CPUs: A Study Based on RISC-V,” in Proceedings of the SC ’23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis. ACM, 2023.
2. M. Perotti et al., “Ara2: Exploring single- and multi-core vector processing with an efficient RVV 1.0 compliant open-source processor,” IEEE Transactions on Computers, vol. 73, no. 7, pp. 1822–1836, 2024.