DATE 25
DESIGN, AUTOMATION AND TEST IN EUROPE
THE EUROPEAN EVENT FOR ELECTRONIC SYSTEM DESIGN & TEST
31 MARCH – 2 APRIL 2025
LYON, FRANCE
CENTRE DE CONGRÈS DE LYON

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

ETH zürich

ALMA MATER STUDIORUM A.D. 1088

ISOLDE RISC-V

European Commission

Chips JU

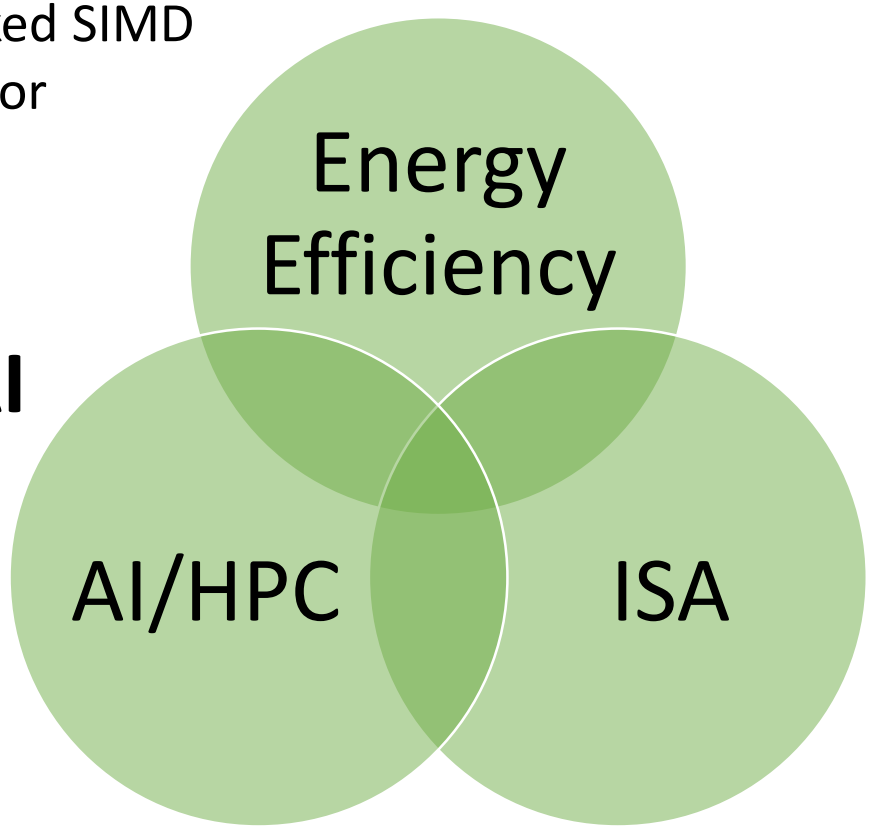
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Case for Vector Processors

Data Parallelism

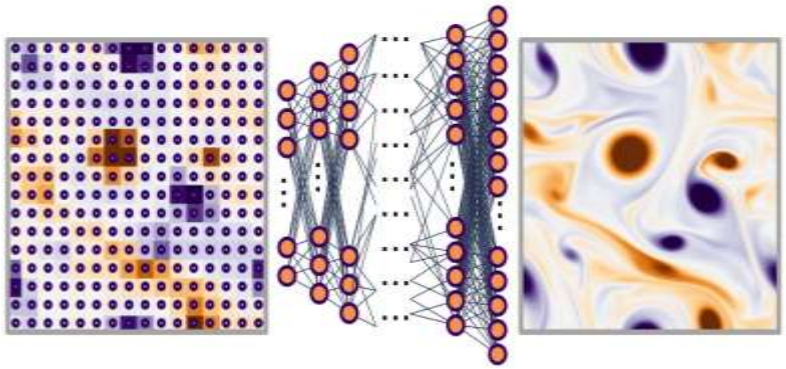
Amortize IF/DEC costs

- Packed SIMD
- Vector
- GPU



Long Vectors in HPC / AI

- Physical Modelling
- Large Language Models

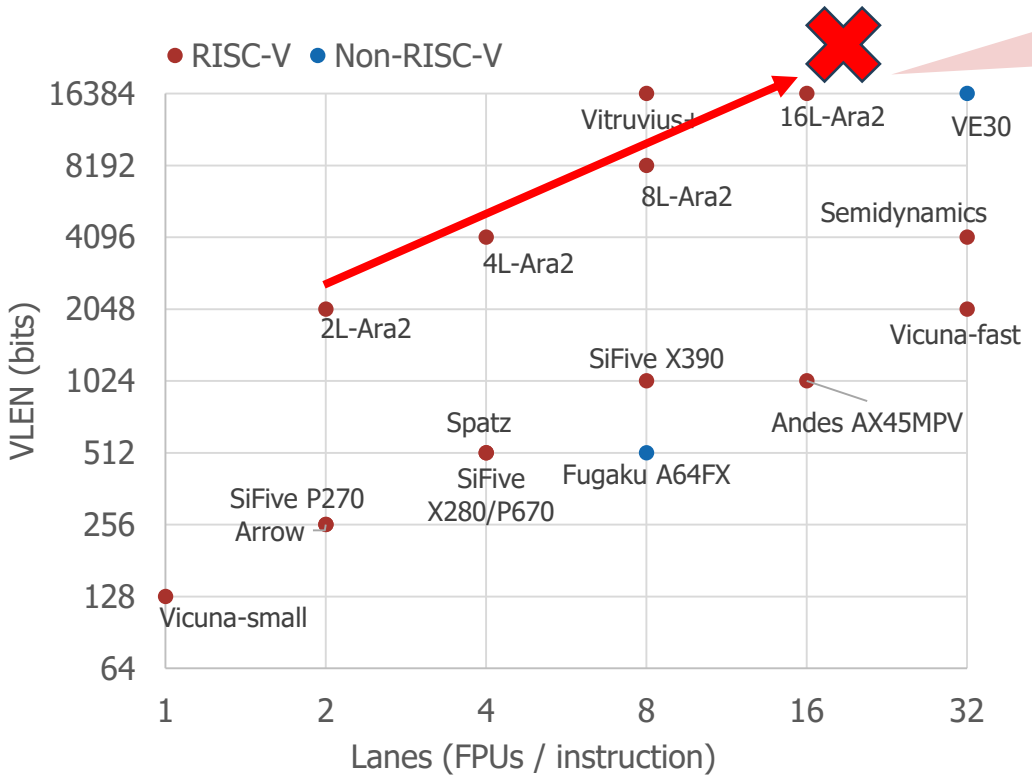


Vector ISA

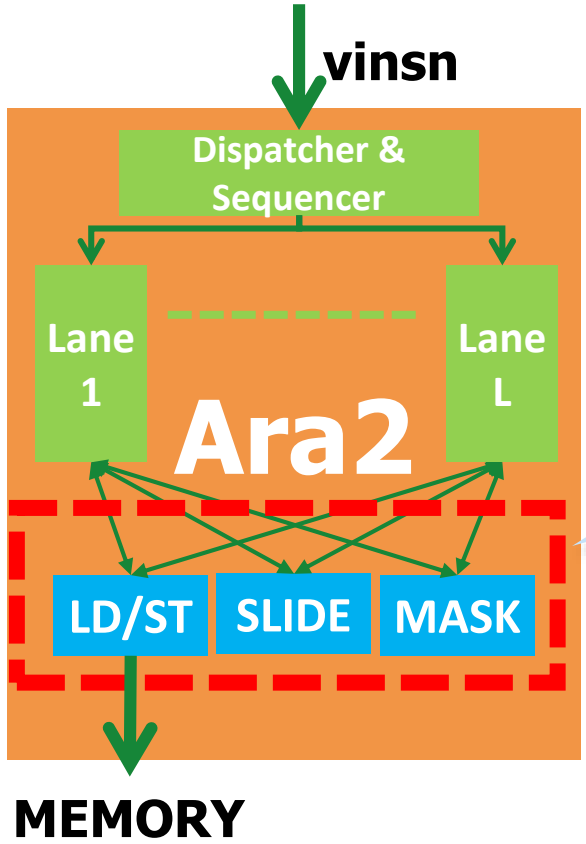
- Arm SVE **(2 Kibits),**
- SX-Aurora **(16 Kibits)**
- RISC-V V **(64 Kibits)**

arm NEC RISC-V

Do Vector Processors scale?



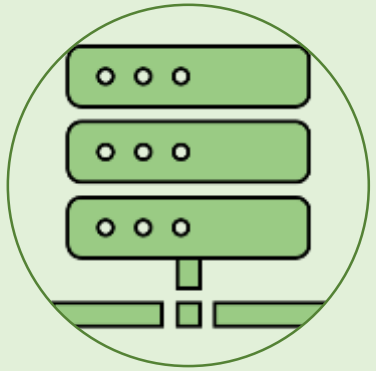
Scales only up to 16 Kibits



All-2-All Interconnect
(L^2 complexity)

AraXL Goals

Scale to **+16 Lanes**



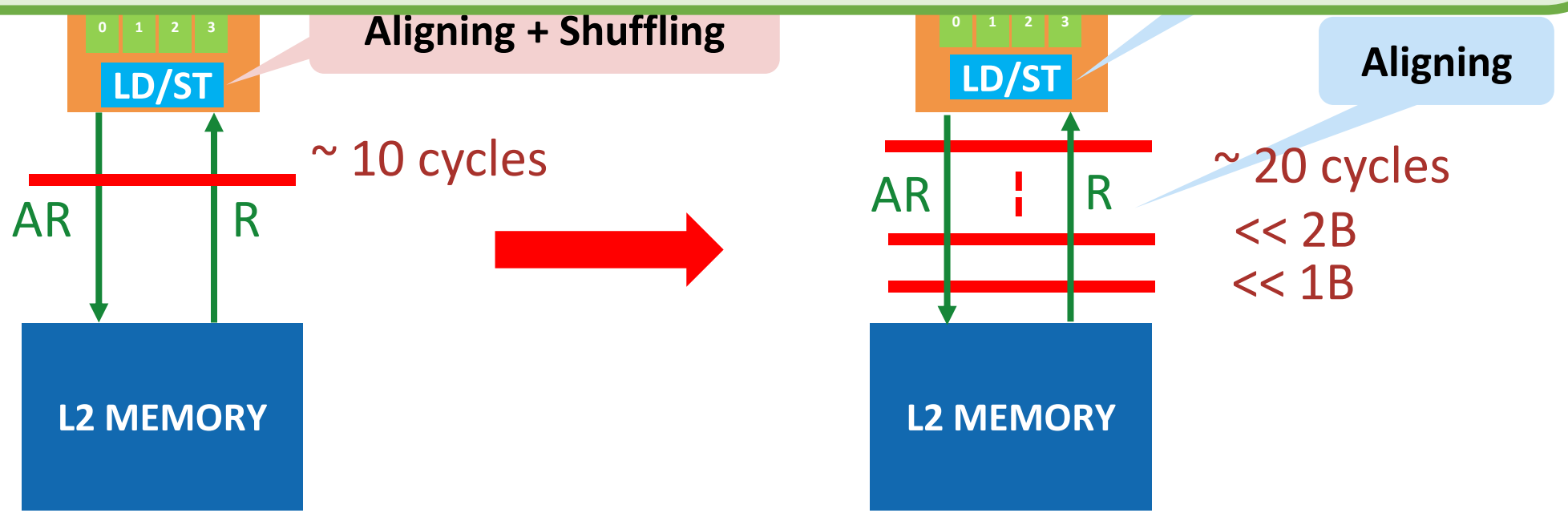
Long vectors in **HPC/AI**

High **Area/Energy Efficiencies**

Solution to Scalability

- Long vectors can tolerate latency [1]

Utilize Latency → Scalable!

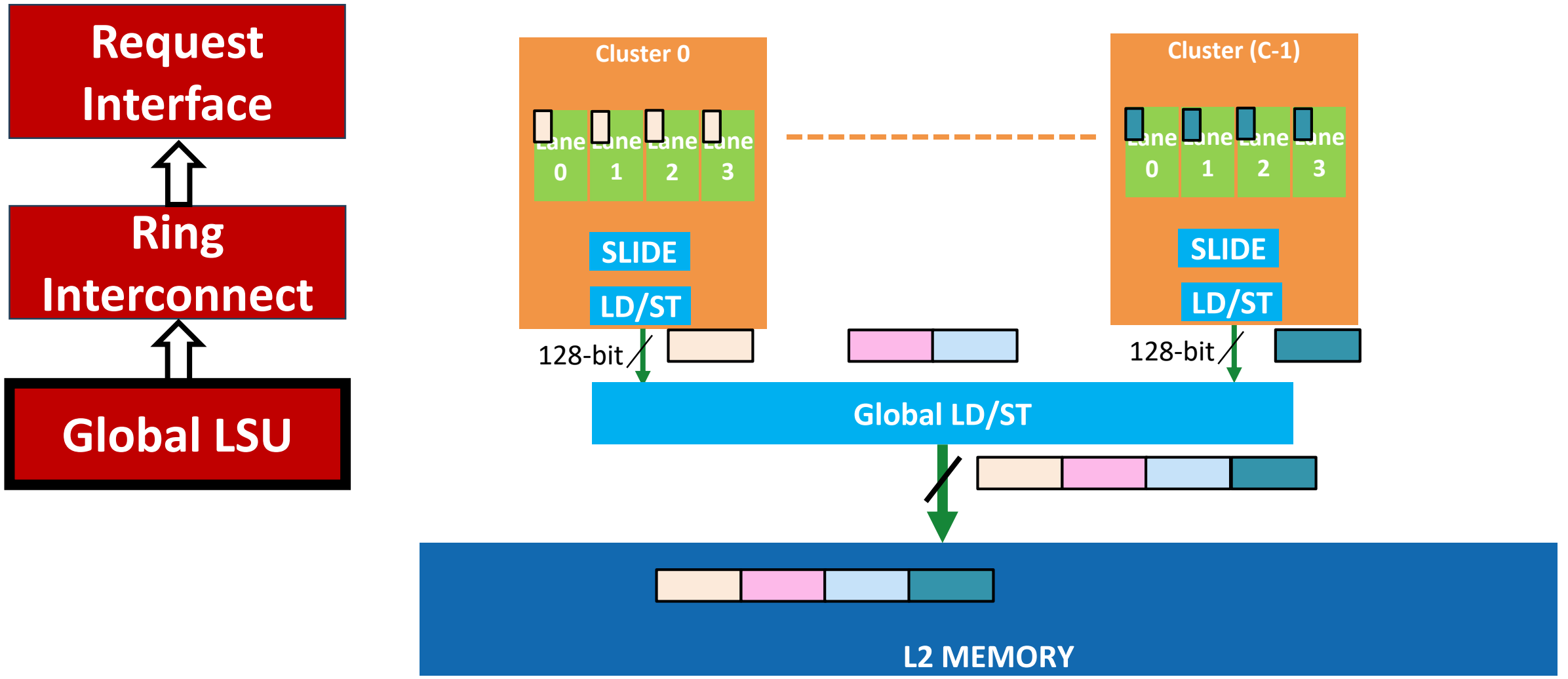


[1] Vizcaino, Pablo, et al. "Short reasons for long vectors in HPC CPUs: a study based on RISC-V." *Proceedings of the SC'23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*. 2023.

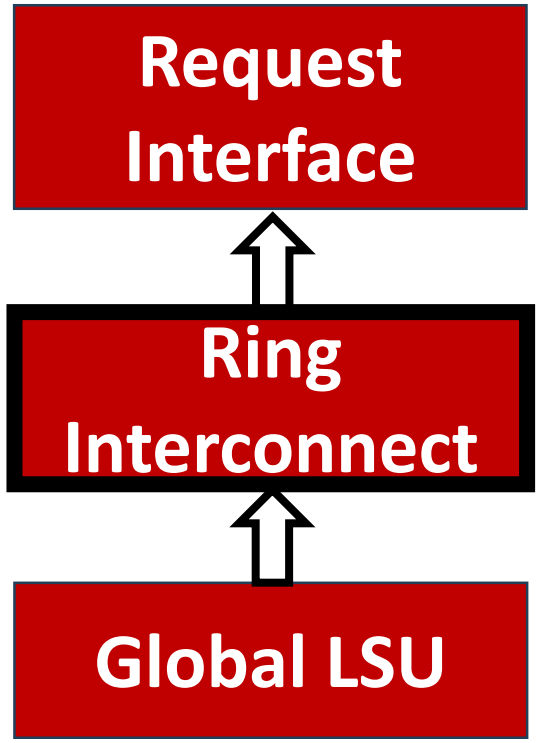
AraXL architecture

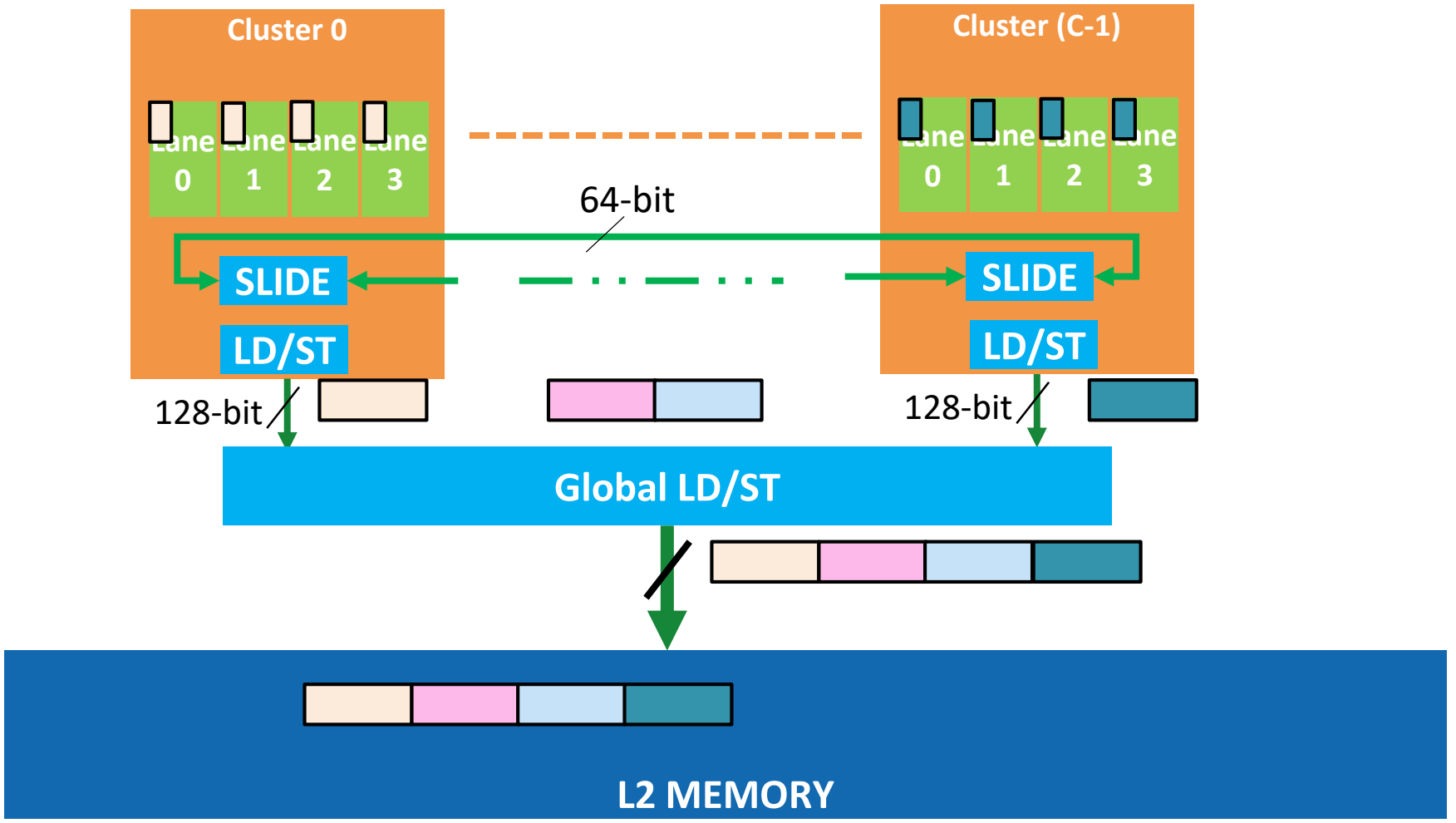
- ✓ **Multiple 4-lane Ara2 instances**
- ✓ **Design pipelined low-cost interconnects**

Global Load-Store Unit

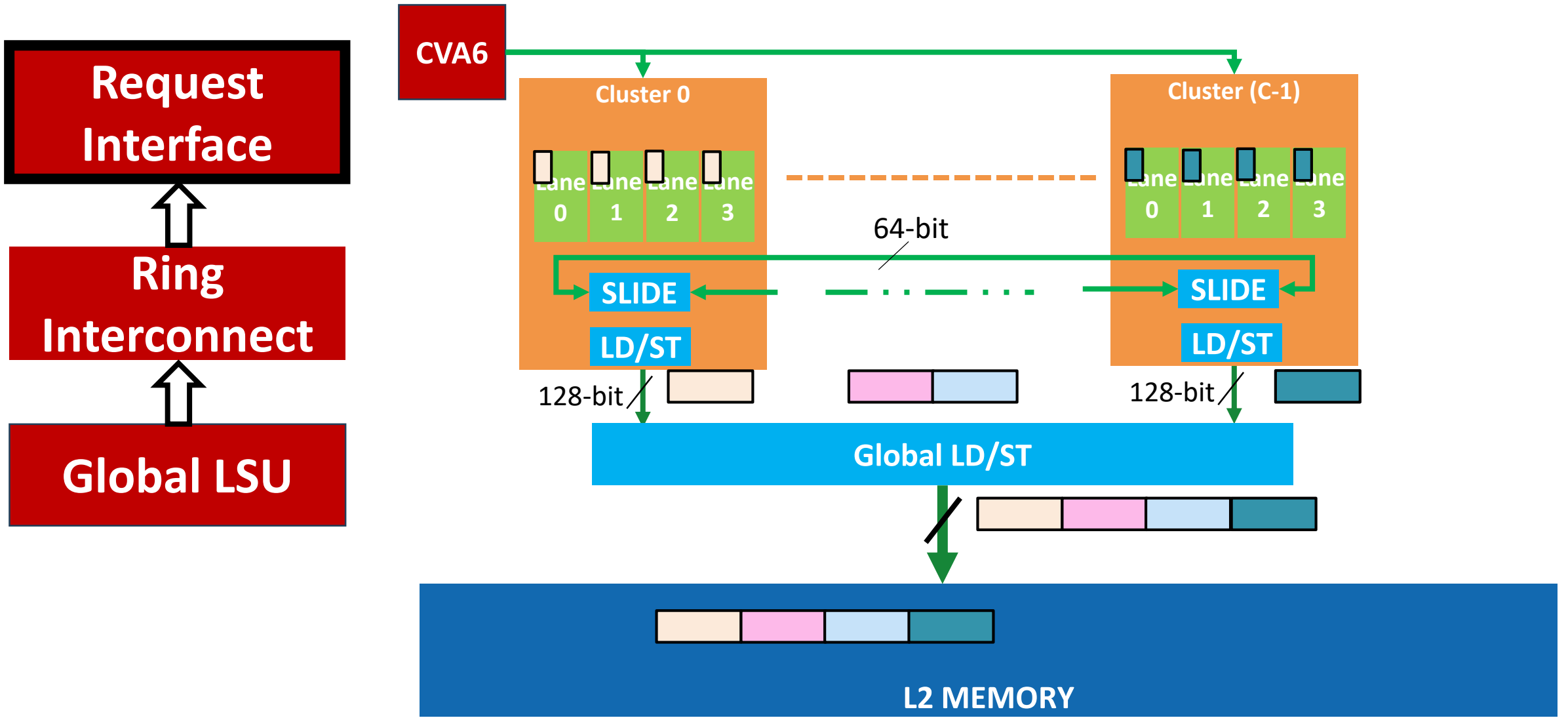


Ring Interconnect



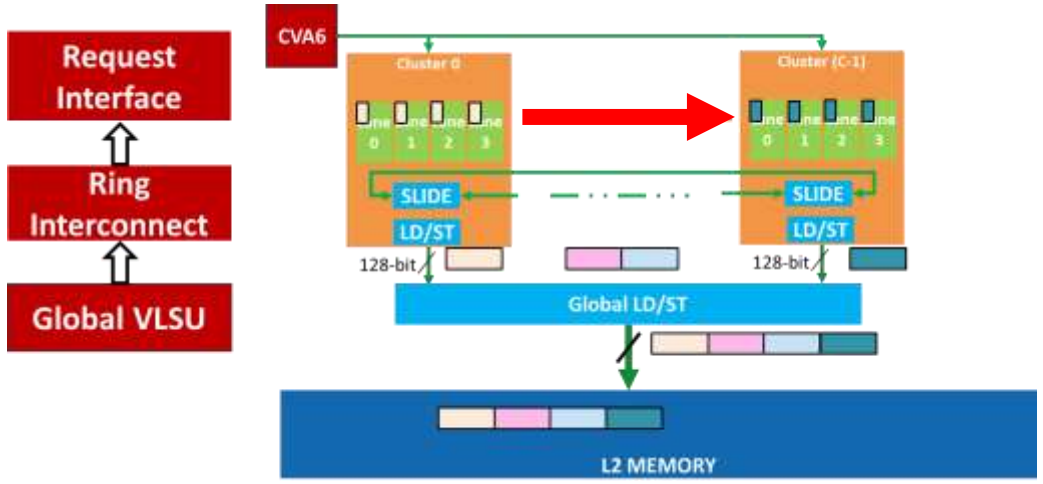


Request Interface



Performance Characteristics

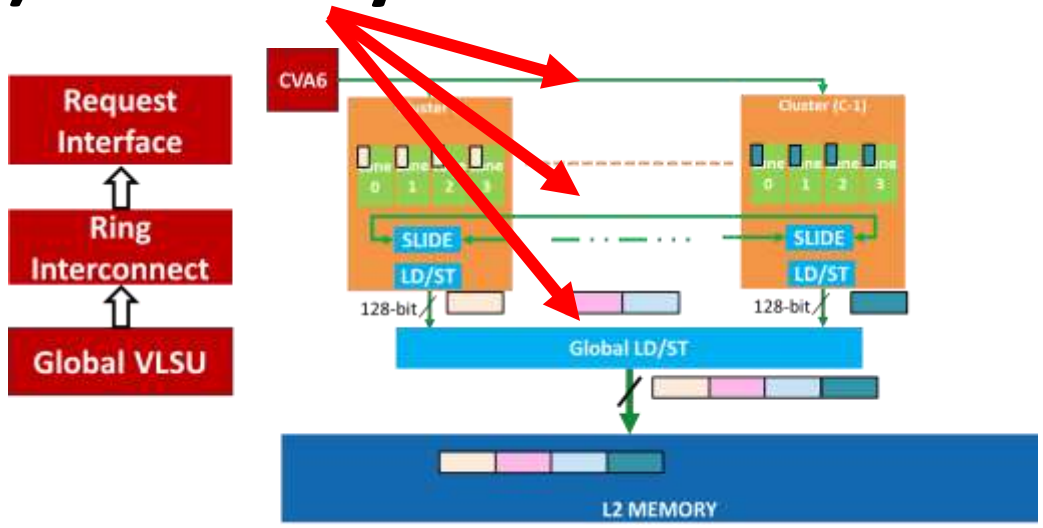
1) ↑ Number of Clusters



Linear Performance scaling

$8 \rightarrow 64$ Lanes $\Longrightarrow$ 8x Performance

2) ↑ Latency of interconnect

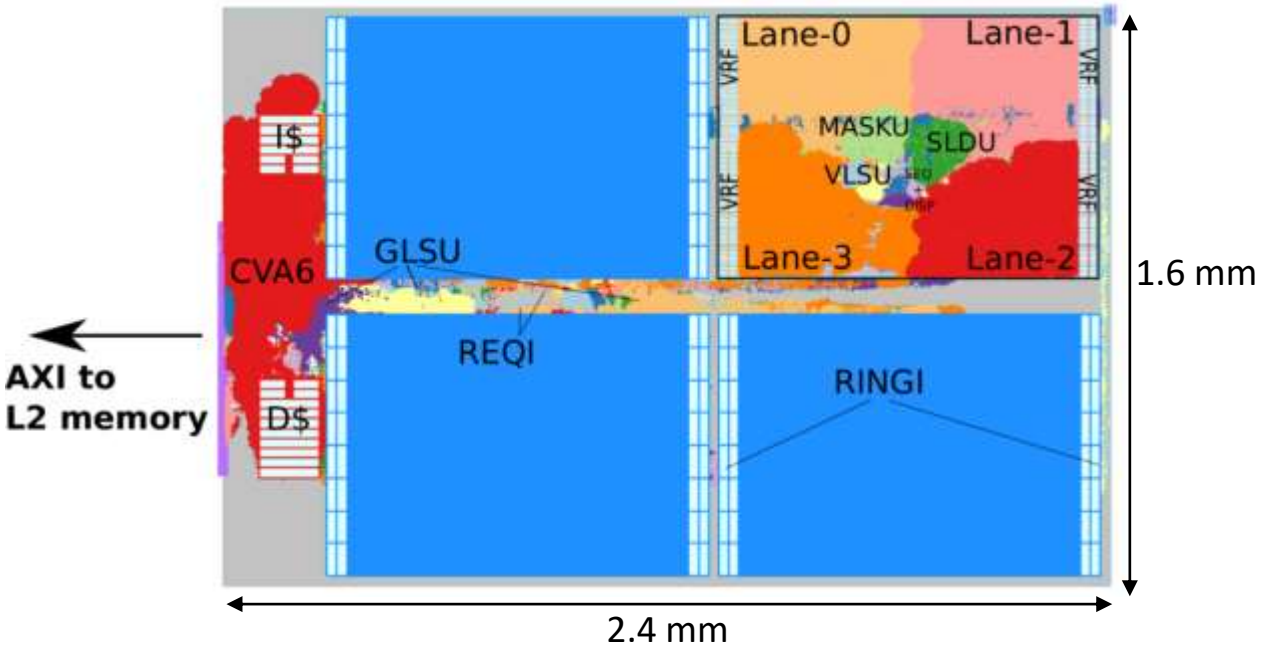


High Latency Tolerance

<2% drop in performance for 64 lane AraXL

Implementation in gf22nm..

Hierarchical Flow



Linear Area Scaling

$Area_{64L} = \mathbf{3.8} \times Area_{16L}$

Frequency Scaling

Maximum frequency of 1.15 GHz (TT/ 0.8V /25C) for 64 lanes

Energy Efficiency

Peak Performance **146 GFLOPS**

Peak Energy efficiency **+40 GFLOPS/W**

We present AraXL

The largest RISC-V vector processor architecture!

Scale to +16 Lanes	**Achieves 64 Lanes** (Highest amongst RISC-V vector processors)
Long vector applications	**Achieves 64 Kibits of VLEN** (Highest among all vector processors)
High Energy Efficiencies	**Achieves peak energy efficiency +40 GFLOPS/W**

Looking forward to meeting you in the poster session…..