TCDM Burst Access: Breaking the Bandwidth Barrier in Shared-L1 RVV Clusters Beyond 1000 FPUs

Diyou Shen[1], Yichao Zhang[1], Marco Bertuletti[1], Luca Benini[1,2]
[1]Integrated Systems Laboratory, ETH Zurich; [2]DEI, University of Bologna

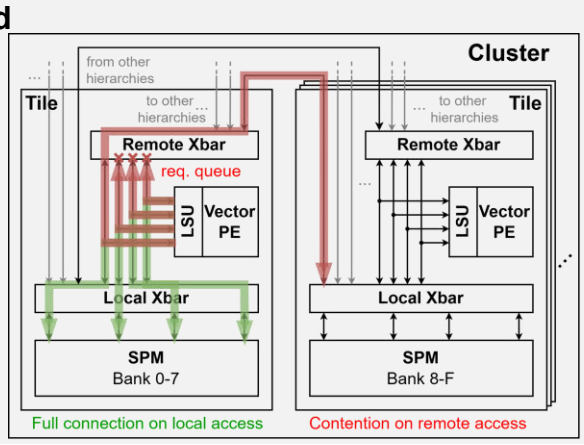
1 Challenge: Scaling Up a Cluster

Larger cluster is preferred

- Better memory utilization
- No data duplication
- Low latency

Challenge to scale up

- Hierarchical design for physical feasibility
- Limit the bandwidth



2 Contributions: TCDM Burst Access

Pack multiple requests from Vector PE into a Burst

- Resolve the request congestion
- Only few extra bits for burst information

Pack multiple responses in a parallel response

- Widen the data field only
- HW Configurable: double (GF2), four times (GF4),…

3 Implementation

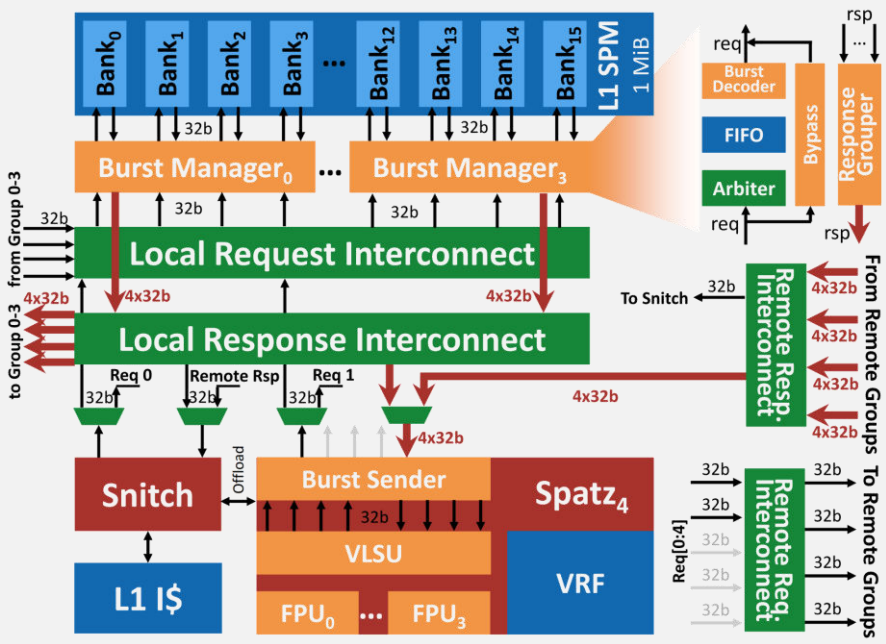
Attach a Burst Sender to Vector PEs

- Collect unit-stride loads from Vector PEs and pack into a burst
- Only add few bits in request for burst information

Add Burst Managers before L1 banks to handle bursts

- Disassemble burst requests into parallel requests
- Pack responses in the widen response data fields

Implement on MemPool-Spatz[1,2] cluster testbeds



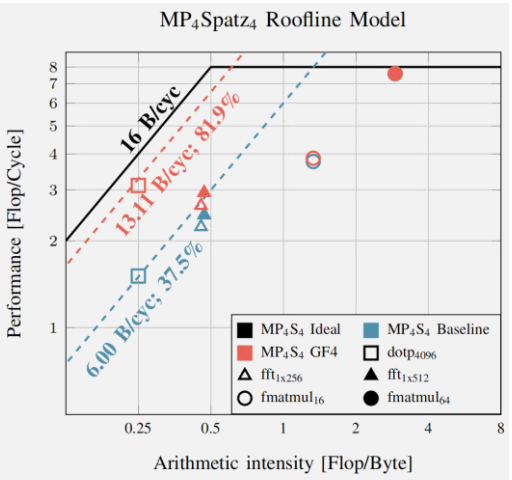
4 Results

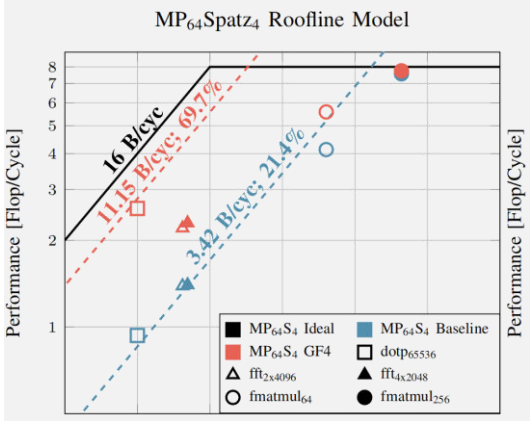
12-nm technology node implementation

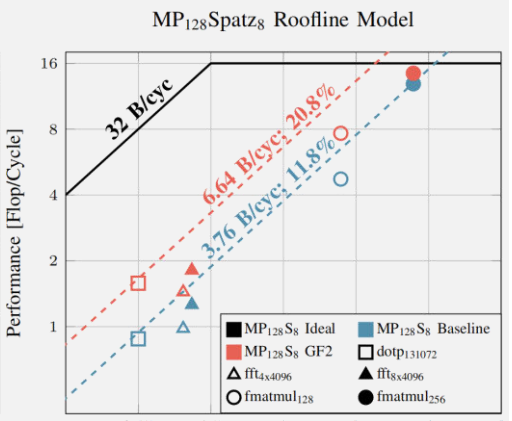
- MemPool-Spatz-based many-core RISC-V vector architecture
- Different cluster scales:
  - **MemPool$_4$Spatz$_4$**: 16-FPU cluster organized in 4 Tiles
  - **MemPool$_{64}$Spatz$_4$**: 256-FPU cluster organized in 64 Tiles
  - **MemPool$_{128}$Spatz$_8$**: 1024-FPU cluster organized in 128 Tiles

Roofline Models

- **Widen the actual BW**
  - **3.26x** on MP64S4
- **Boost real-world kernels**
  - Up to **2.76x** on memory-bound kernels
  - More than **1.3x** improves on matmul

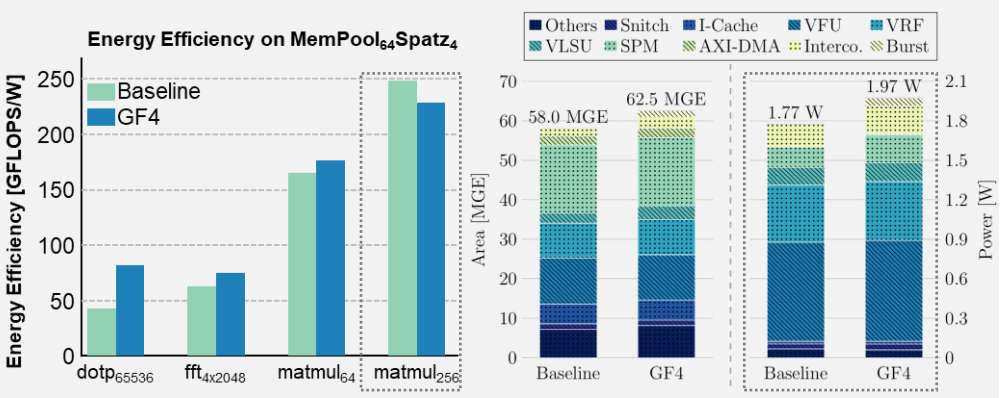






Area and Power

- No maximum frequency degradation, **fully routable**
- **< 8%** area overhead for GF4 on MemPool$_{64}$Spatz$_4$
- **Energy efficiency boost** for memory-bound kernels



References

1. Samuel Riedel, Matheus Cavalcante, Renzo Andri, and Luca Benini. 2023. MemPool: A Scalable Manycore Architecture With a Low-Latency Shared L1 Memory. IEEE Trans. Comput. 72, 12 (Dec. 2023), 3561–3575.
2. Matheus Cavalcante, Domenic Wüthrich, Matteo Perotti, Samuel Riedel, and Luca Benini. 2022. Spatz: A Compact Vector Processing Unit for High-Performance and Energy-Efficient Shared-L1 Clusters. In Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design (ICCAD '22). Association for Computing Machinery, New York, NY, USA, Article 22, 1–9.

GitHub