MX: Ultra-Low Overhead, Energy-Efficient Matrix Multiplication on RISC-V Vector ISA

Matteo Perotti[1], Yichao Zhang[1], Matheus Cavalcante[1], Enis Mustafa[1], Luca Benini[1,2]
[1]ETH Zurich, Switzerland; [2]University of Bologna, Italy

1 Introduction

Matrix Multiplication is crucial across all the computing domains

Its performance and power consumption influence the whole system
Optimizing it is paramount from the servers to the embedded devices

Where is Matrix Multiplication run?

- Domain-specific accelerators (fast and efficient, but algorithm-specific)
- General-purpose processors (slower, but flexible and easy to program)

RISC-V is an **open-source ISA** for general-purpose processors.
Its **vector extension** helps speed up matrix operations, but its vector register file is **energy hungry**! How to reduce the number of accesses?

2 Contributions

MX – ISA Extension to RISC-V V

- Boost Matrix Multiplication energy-efficiency at negligible area cost

MX – PPA, Energy Efficiency analysis

- Implement MX in Dual- and 64-core vector architecture in 12-nm technology

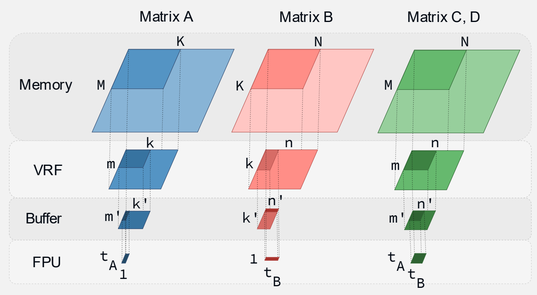
3 Implementation

Give the vector processor matrix capabilities

The load-store unit can handle matrices (matrix load/store)

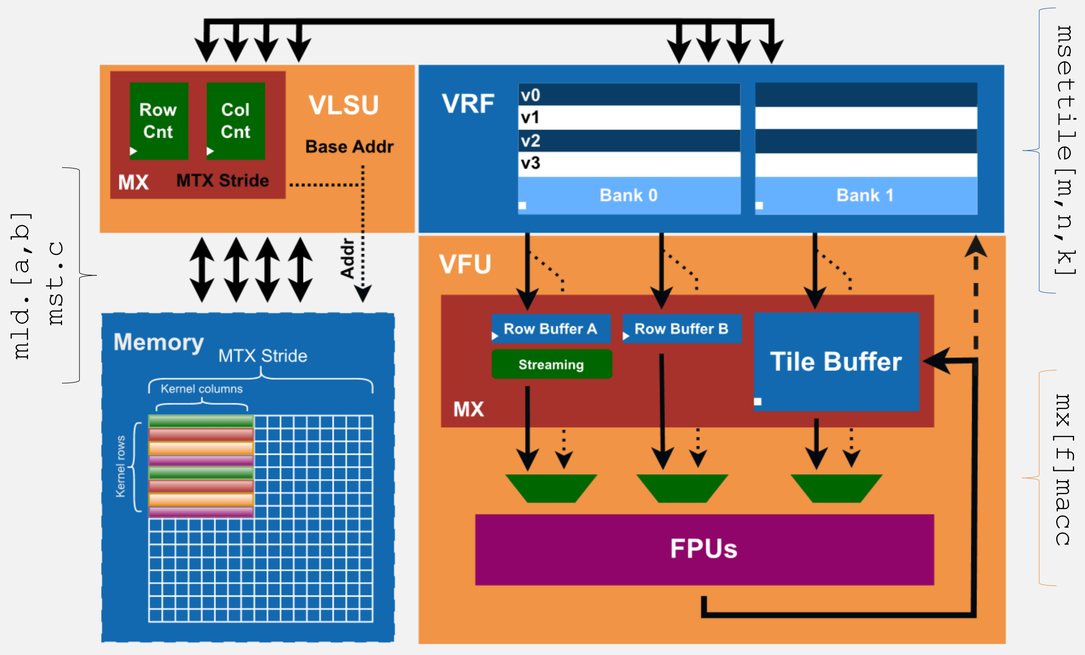
Cut down the Vector Register File (VRF) accesses

- Introduce a small latch-based buffer close to the FPUs
- Accumulate a whole matrix tile inside the buffer (outer product)
- Write-back to the VRF when the accumulation is over



Re-use all the RISC-V vector functional units

Additions: tile buffer, broadcasting engine, control in the load-store unit



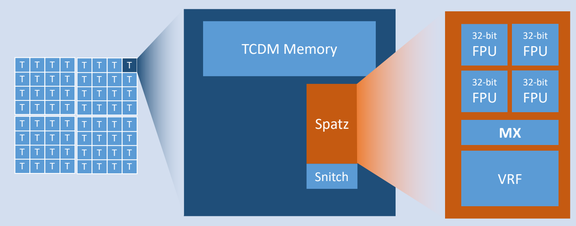
4 Results and Discussion

12-nm technology node implementation

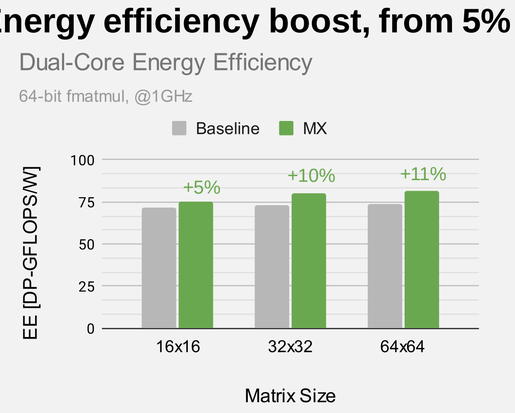
- Dual-core (64-bit) and 64-core (32-bit) cluster
- Spatz[1]-based multi-core RISC-V vector architecture
- MX support in each Spatz

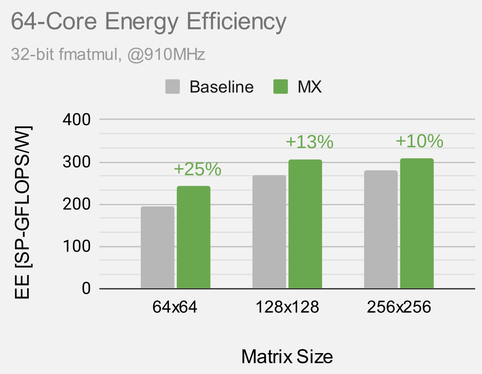
64-core cluster architecture

- Mempool[2]-Spatz based
- 64 tiles (see figure)

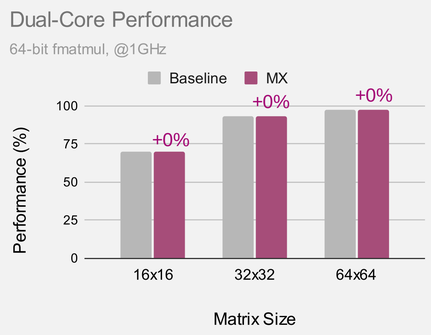


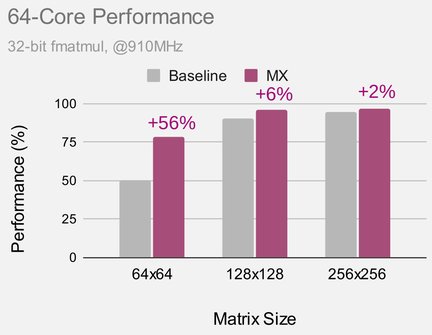
Energy efficiency boost, from 5% to 25%



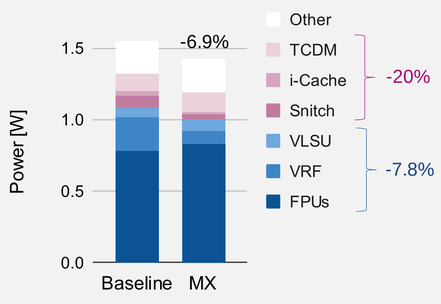


No maximum frequency degradation. IPC better or equal

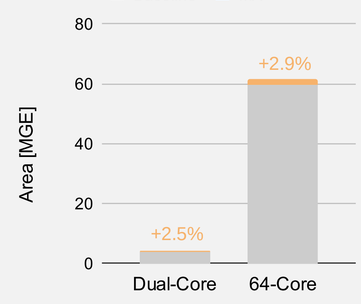




VRF-related power drops



Cost? < 3% area overhead!



5 Conclusion

MX is an ISA matrix extension for RISC-V V

It boost matrix multiplication's **energy efficiency** up to +**25%**, with up to **1.5× speedup** for **less than 3% area** cost, by adding a small close-FPU tile buffer to increase the matrix data reuse and save energy.

6 References

[1] Matheus Cavalcante, et Al., "Spatz: A Compact Vector Processing Unit for High-Performance and Energy-Efficient Shared-L1 Clusters". ICCAD '22.

[2] S. Riedel, et Al., "MemPool: A scalable manycore architecture with a low-latency shared L1 memory," IEEE Transactions on Computers, 2023.