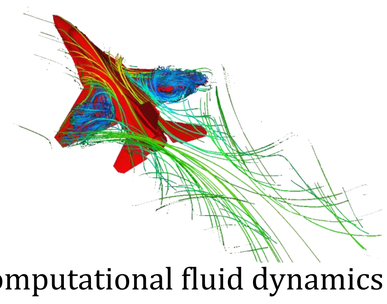
VRP: a Variable Precision Accelerator for Scientific Computing Applications

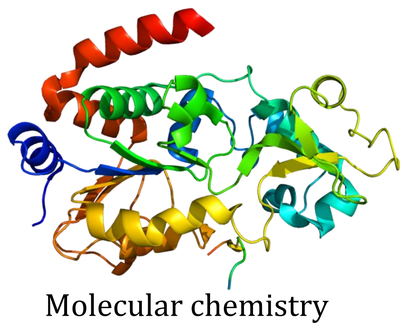
Andrea Bocco, Eric Guthmuller, Jerome Fereyre, and César Fuguet (CEA/LIST/DSCIN/LSTA)

Context

Scientific computing applications



Computational fluid dynamics



Molecular chemistry

High Performance Computing (HPC)

- Problem size ↑
- Rounding and Quantization errors ↑
- Convergence speed ↓

WARNING: Result may not converge!

1: Increase computing precision

2: Reduce rounding error

3: Increase convergence speed → *Speedup execution*; *Enable kernel convergence*

State of the Art

Used kernels :

- Conjugate Gradient (CG)
- Preconditioned Conjugate Gradient (PCG)
- BiConjugate Gradient (BiCG)

SoA = software based solutions :

- MPFR library [1], libquadmath
  - \+ Offer scalability
  - − Execution time overhead w.r.t hardware accelerator

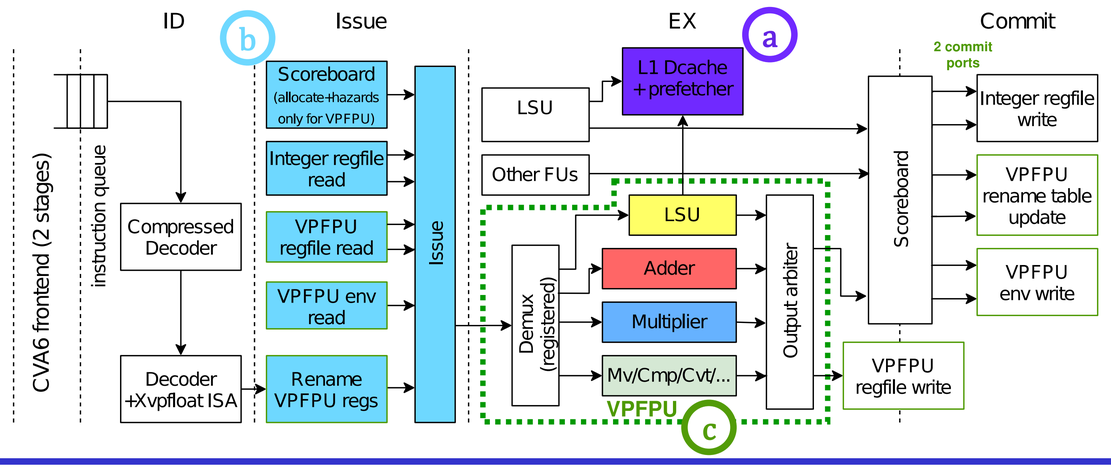
➢ An hardware accelerator is needed

Our contribution

The modified CVA6 [2] with embedded Variable-Precision FPU and Variable-Precision register file

(a) New L1 cache [3] to mitigate cache miss latency

(b) Modified issue stage accommodating extended precision floating-point data up to 512-bit

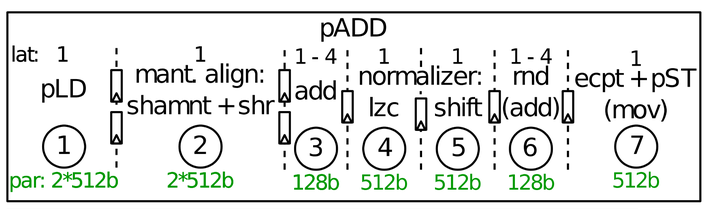


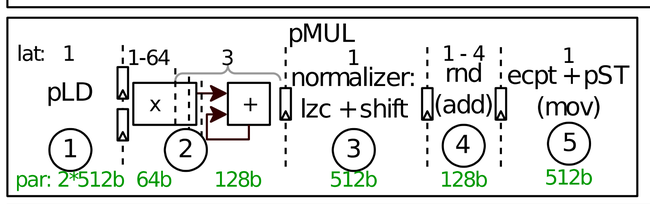
(c) Adapted execute stage housing the VPFPU

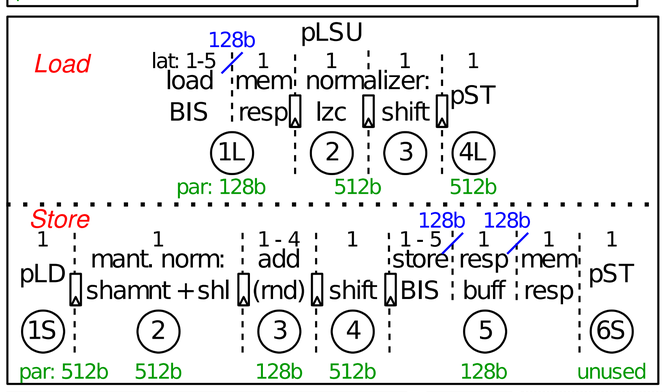
Variable Precision Floating Point Unit (VPFPU)

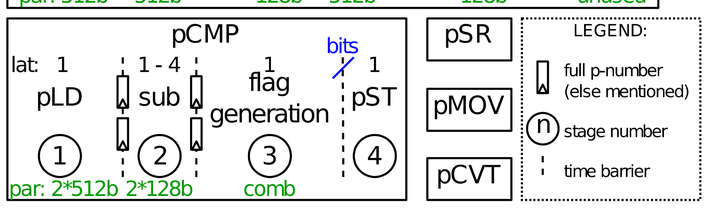
The VPFPU is designed to match the CVA6 operating frequency[2,4]

- Has an heterogeneous pipeline (par) on 64, 128, and 512 bits
- Computing precision, memory data and exponent bit-width, are programmable at run-time









Load Store Unit (pLSU)

- Handles operations on 128-bit data chunks
- Supports mis-aligned accesses
- Resolve hazards (read-after-write and write-after-read)

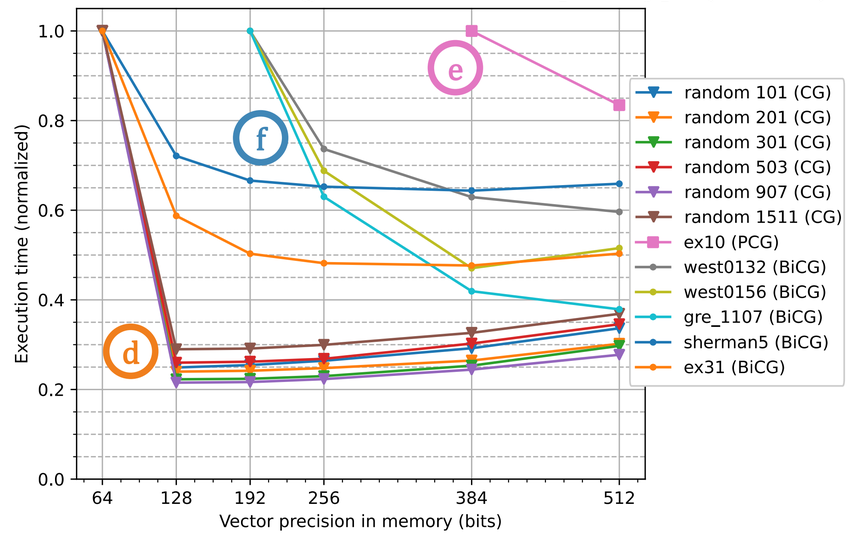
Benchmarking

Experiments evaluating CG, PCG, and BiCG solvers across different matrix sets, random and from the Florida Sparse Matrix Collection, executed on a single VRP core FPGA prototype (91 MHz)

(d) 79% execution time decrease

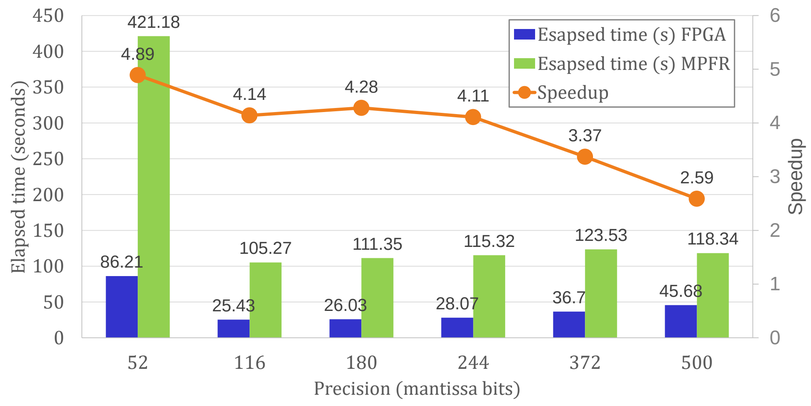
(e) Some matrices only converge with extended precision

(f) BiCG: convergence execution time is optimum between 192 and 384 bits



VRP vs. Intel(R) Xeon(R) E3 1240v5

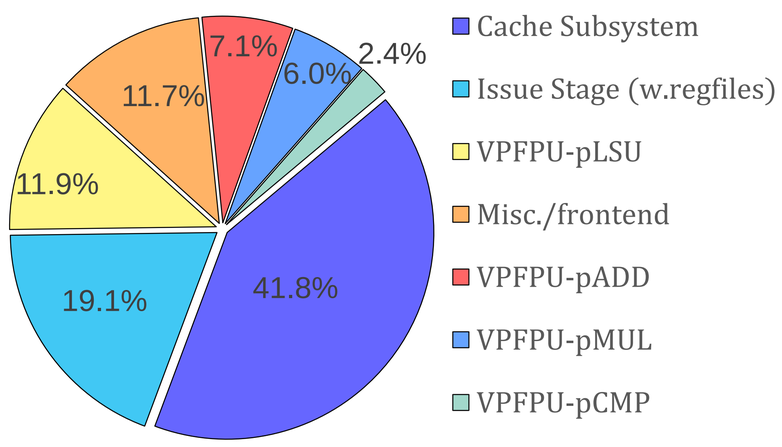
The VRP (91 MHz) on CG with random 1511 matrix runs between 2.5x and 5x faster w.r.t. MPFR running on a single thread on an Intel(R) Xeon(R) E3 1240v5 CPU (3.50GHz).



Better execution time (up to x5) while running at a 38x slower clock frequency !

ASIC implementations reach higher frequencies with smaller area, enabling many-core VRP architectures to provide hundreds of GFLOPS at high precision, limited only by memory bandwidth.

Area distribution



Future work

1) Implement Out Of Order hardware support → 2) Sparse dataset optimization → 3) Implement Eigen solvers support

[1] Laurent Fousse et al. “MPFR: A Multiple-Precision Binary Floating-Point Library with Correct Rounding”. In: ACM Trans. Math. Softw. (2007). DOI: 10.1145/1236463.1236468.
[2] Yves Durand et al. “Accelerating Variants of the Conjugate Gradient with the Variable Precision Processor”. In ARITH (2022). DOI: 10.1109/ARITH54963.2022.00017.
[3] César Fuguet. “HPDcache: Open-Source High-Performance L1 Data Cache for RISC-V Cores”. ACM International Conference on Computing Frontiers. (2023), DOI: 10.1145/3587135.3591413.
[4] Eric Guthmuller et al., "Xvpfloat: RISC-V ISA Extension for Variable Extended Precision Floating Point Computation," in IEEE Transactions on Computers (2024), DOI: 10.1109/TC.2024.3383964