Enabling mixed-precision with VerifiCarlo: Sharing CEEC experience

Roman Iakymchuk[1]
jointly with Yanxiang Chen & Gülçin Gedik (UmU), and
Pablo de Oliveira Castro (Paris-Saclay, UVSQ)

Umeå University and Uppsala University, Sweden
riakymch@cs.umu.se

CEEC Webinar
March 04, 2025

UMEÅ UNIVERSITY

CEEC Centre of Excellence in Exascale CFD

[1] This research is partially funded by EuroHPC JU CoE CEEC (No. 101093393)

Outline

Energy efficiency: from hardware to NLA

- Supercomputing is constrained by **power consumption**

→ Power-efficient hardware
  - RIKEN's Fugaku w A64FX (FP64:FP32:FP16 = 1:2:4)
  - EPI (ARM, FPGA, RISC-V)
  - **Jülich to host the first EPI-based supercomputer**

- Numerical linear algebra is known to be dominant by **double precision**

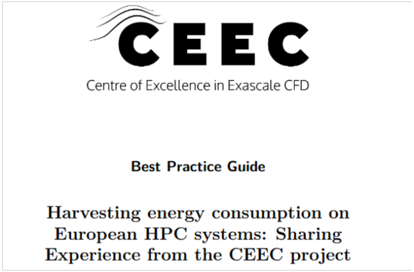
→ **Energy-efficient algorithms**

math Mixed and adaptive precision computing

code Communication hiding or avoiding

tools Numerical abnormalities and precision cropping

Measuring energy consumption


CEEC

Centre of Excellence in Exascale CFD

Best Practice Guide

Harvesting energy consumption on European HPC systems: Sharing Experience from the CEEC project

Iakymchuk et al. Zenodo, 2024
doi:10.5281/zenodo.13306639

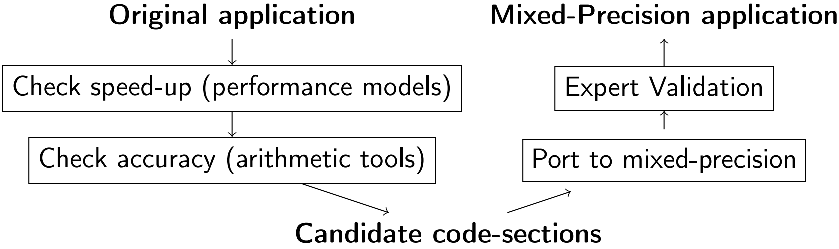
- **More complex than measuring time-to-solution**
- Measurements require elevated privileges

Objectives

- Facilitate energy measurements on the European HPC systems
- Teach the community how to conduct such measurements
- Provide examples with easy-to-use guide

Methodology

Methodology to enable mixed-precision algorithmic solutions in applications with accuracy guarantees.



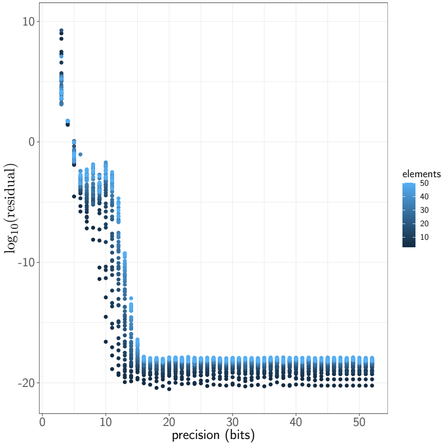
Nekbone w Vprec

$$\boxed{Ax = b}$$

while $(\tau > \tau_{\max})$

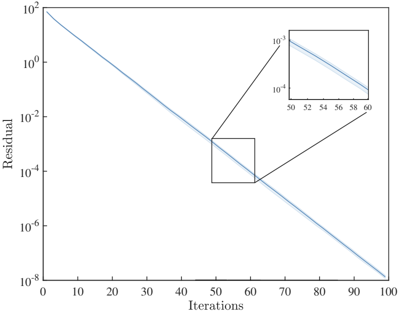
Step	Operation
$S1:$	$w := Ad$
$S2:$	$\rho := \beta / < d, w >$
$S3:$	$x := x + \rho d$
$S4:$	$r := r - \rho w$
$S5:$	$z := M^{-1} r$
$S6:$	$\beta := < z, r >$
$S7:$	$d := (\beta / \beta_{old}) d + z$
$S8:$	$\tau := < r, r >$

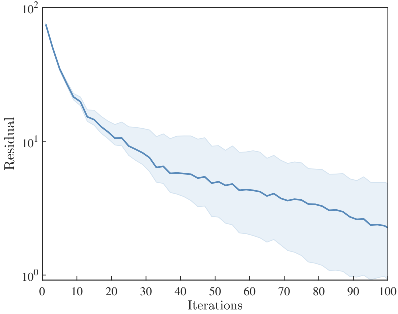
end while



Nekbone w MCA for FP32

Entire program, no preconditioner

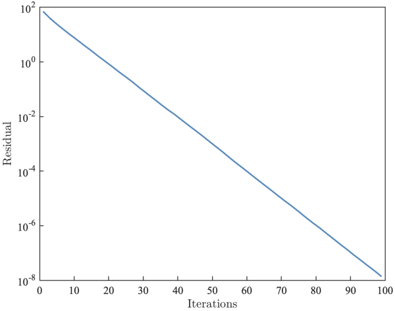


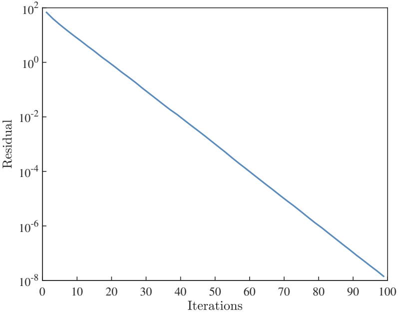


- Random Rounding (rr) mode (left)
- MCA (mca) mode (right)
- Issue in initialization $10^9 \cos(x) \rightarrow$ focus on the solver only

Nekbone w MCA for FP32

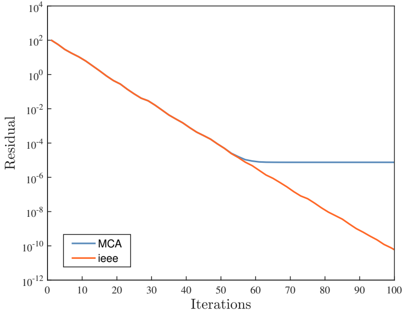
Only the CG loop, no preconditioner

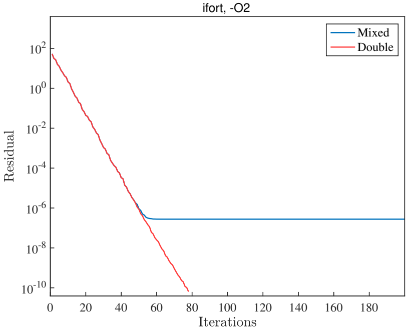




- Random Rounding (rr) mode (left)
- MCA (mca) mode (right)

Nekbone: CG with preconditioner

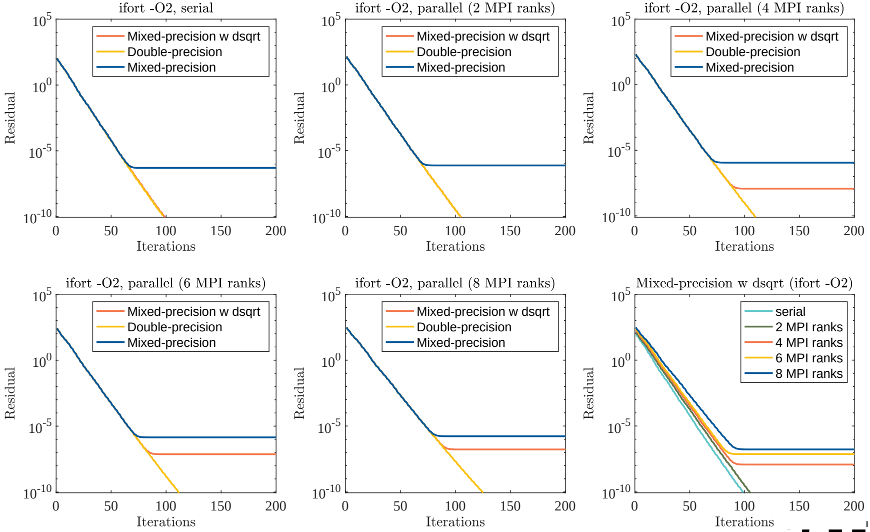




- Verificarlo predicts stagnation (left plot)
- Stagnation in the mixed-precision Nekbone (right plot)
- Stagnation occurs after 61st iteration w residual $r = 7.94 \times 10^{-6}$

Nekbone: CG with preconditioner

Square root



Nekbone: CG with preconditioner

Mixed-precision strategies

Mixed-precision strategy evaluation

GSO mode	MPI	Precond. Ops	Precond. GSO	PCG Ops	PCG GSO	Conv.
pairwise	< 4	fp32	fp32	fp32	fp32	stagnates
		fp64	fp64			converges
		fp32+fp64sqrt	fp32			converges
		fp32	fp64			converges
	≥ 4	fp64	fp64		fp32	stagnates
		fp32+fp64sqrt	fp32			stagnates
		fp32	fp64			stagnates
		fp32+fp64sqrt	fp64		fp64	converges
		fp64	fp64		fp64	converges

- Gather-Scatter Operations have strong impact on convergence
- dot product impacts convergence too

Nekbone: CG with preconditioner

128 elements per MPI rank on MareNostrum 5: Intel Sapphire Rapids 8460Y+ w 40 cores

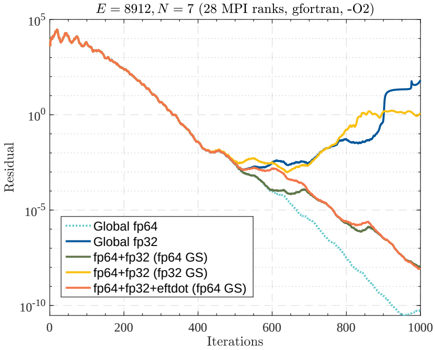
MPI ranks/ secs	8	20	40	80
Double	5.79	8.98	13.33	24.02
Mixed-1[a]	4.79	5.99	9.42	14.85
Gain-1	17.27%	33.30%	29.33%	38.18%
Mixed-2[b]	4.88	6.02	8.37	10.11
Gain-2	15.72%	32.96%	**37.21%**	**2.38x**

MPI ranks/ joules	8	20	40	80
Double	1875	3035	7191	16261
Mixed-1[a]	939	2088	3133	7428
Gain-1	1.97x	31.20%	2.30x	2.19x
Mixed-2[b]	934	1566	2570	3482
Gain-2	2.01x	1.94x	**2.80x**	**4.67x**

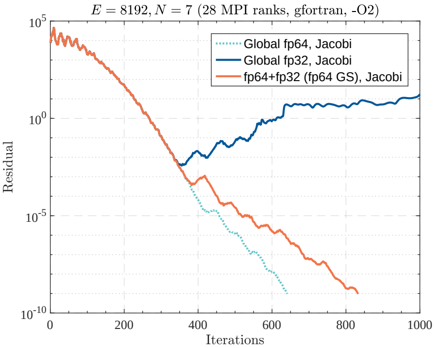
[a] fp64+fp32+fp64sqrt (fp32 GS, allreduce)
[b] fp64+fp32+fp64sqrt (fp64 GS, pairwise)

CEEC

Neko: convergence



CG with the identity preconditioner



CG with the Jacobi preconditioner

- Solving the Poisson's equation with Neko
- The winning strategy is fp64+fp32 (fp64 GS)

Neko: CG with the identity preconditioner

16,384 elements with pol. degree 7 on MareNostrum 5

MPI ranks/ secs	8	20	40	80
Double	14.10	6.61	4.80	1.87
Mixed	12.50	5.77	3.41	1.55
Gain	11.35%	12.71%	**28.96%**	**17.11%**

MPI ranks/ joules	20	40	80
Double	14437	11686	10612
Mixed	13403	10509	8033
Gain	7.16%	**10.07%**	**24.30%**

- Solving the Poisson's equation with Neko
- Time is for the CG loop
- Energy measurements are for the entire code

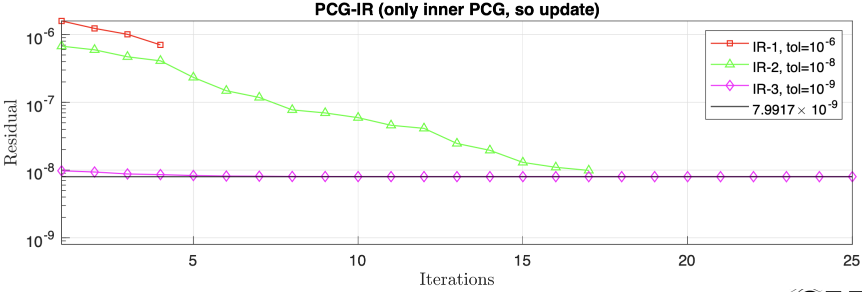
CEEC

Work in progress: PCG with iterative refinement

```
PCG to solve Ax_m = b until tol = 10^-4          ▷ FP32. PCG breaks at 10^-6
for i ← 1, 2, 3 do                               ▷ Run for few iterations
    r = b − Ax_m                                 ▷ FP64
    r_new = r/‖r‖                                ▷ FP64
    PCG to solve Ad_new = r_new until tol = 10^-6, 10^-8   ▷ FP32
    d = d_new ∗ ‖r‖                              ▷ FP64
    x_m = x_m + d                                ▷ FP64
end for
```

Work in progress: PCG with iterative refinement

```
PCG to solve Ax_m = b until tol = 10^-4            ▷ FP32. PCG breaks at 10^-6
for i ← 1, 2, 3 do                                 ▷ Run for few iterations
    r = b − Ax_m                                   ▷ FP64
    r_new = r/‖r‖                                  ▷ FP64
    PCG to solve Ad_new = r_new until tol = 10^-6, 10^-8   ▷ FP32
    d = d_new ∗ ‖r‖                                ▷ FP64
    x_m = x_m + d                                  ▷ FP64
end for
```



Summary

- Computer arithmetic operates with finite precisions
- Use computer arithmetic tools to
  - detect cancellations
  - get the right FP format
  - verify sensitivity of reduced precision
- Enabling mixed-precision in CFD codes:
  - use tools: Verificarlo, gprof, Intel Advisor
  - target the most time-consuming parts
  - reduced time-to-solution by up to 40 % and **energy-to-solution by up to 2x** on MareNostrum5 & LUMI

Summary

- Computer arithmetic operates with finite precisions
- Use computer arithmetic tools to
  - detect cancellations
  - get the right FP format
  - verify sensitivity of reduced precision
- Enabling mixed-precision in CFD codes:
  - use tools: Verificarlo, gprof, Intel Advisor
  - target the most time-consuming parts
  - reduced time-to-solution by up to 40 % and **energy-to-solution by up to 2x** on MareNostrum5 & LUMI

Thank you for your attention !

This research is partially supported by EuroHPC JU CoE CEEC (No. 101093393)

CEEC

References

- Iakymchuk et al. *Best Practice Guide – Harvesting energy consumption on European HPC systems: Sharing Experience from the CEEC project*. Zenodo, Aug 2024
- Chen et al. *Enabling mixed-precision with the help of tools: A Nekbone case study*. In proceedings of PPAM 24. arXiv:2405.11065
- Goldberg. *What every computer scientist should know about floating-point arithmetic*. ACM Comp Sur. Vol 23, No 1. 1991
- Boldo, Jeannerod, Melquiond, Muller. *Floating-point arithmetic*. Acta Numerica, 2023
- Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002
- Butari et al. *Mixed-precision iterative refinement techniques for the solution of dense linear systems*. IJPHCA 2007