Harnessing Reduced Precision for Accurate and Efficient Scientific Computing in HPC

Nima Sahraneshin Samani

Supervisors
Sandra Catalan
José R. Herrero
José Ignacio Aliaga

Fosdem 2025
2nd Feb, 2025

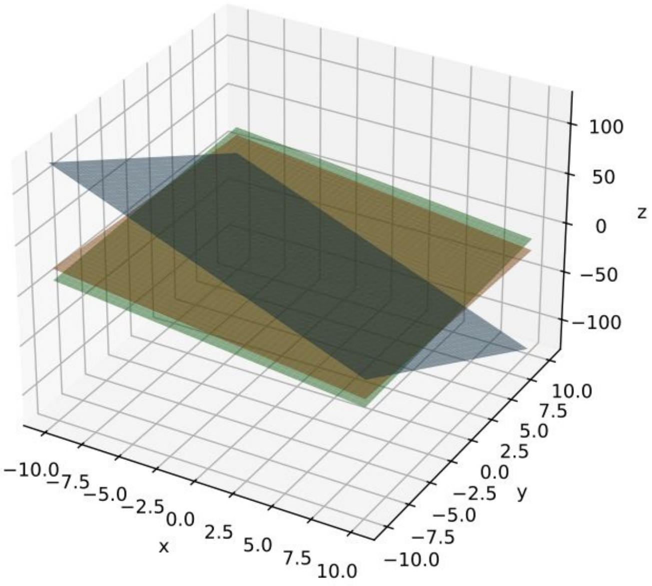
UNIVERSITAT JAUME·I

Motivation

Many scientific and engineering applications rely on heavy **linear algebra calculations**.

Floating-point arithmetic with 64 bits is often used to ensure accuracy.

There's growing interest in using lower-precision formats,



Mixed-Precision Solvers

Idea: Solve **Ax=b** first in FP16 by **LU decomposition**, then refine in FP64 [1].

LU decomposition: It's a way of breaking down a matrix into a lower triangular matrix **(L)** and an upper triangular matrix **(U)**, simplifying the solution of linear systems.

Expected Benefit:

- FP16 is faster and reduces memory movement
- FP64 refinement ensures accuracy

But... does the FP16 solution really help?

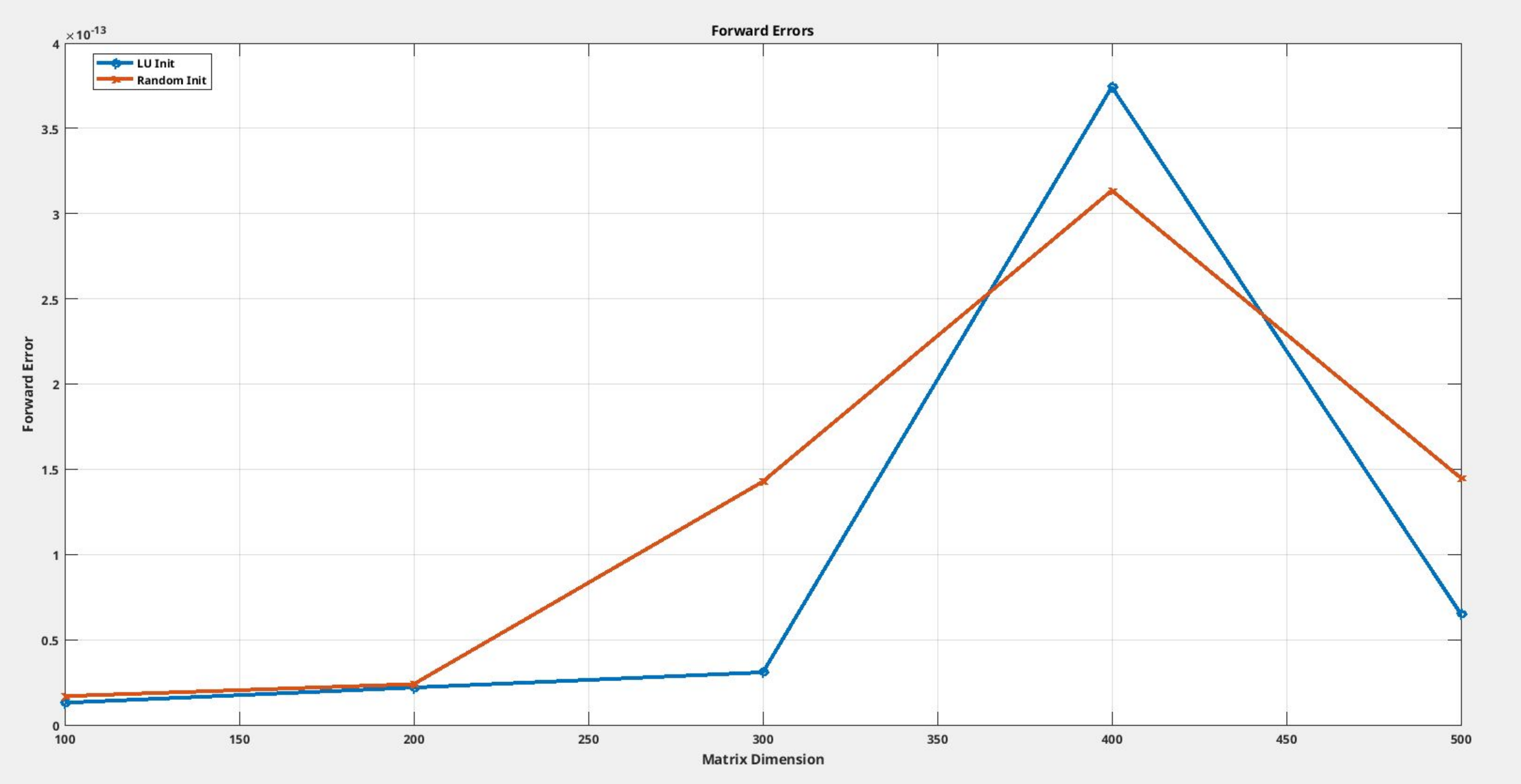
GMRES Convergence: Comparing FP16 LU Initialization vs. Random Start.

experiment: Compare GMRES with and without an initial FP16 LU solution.

Observation:

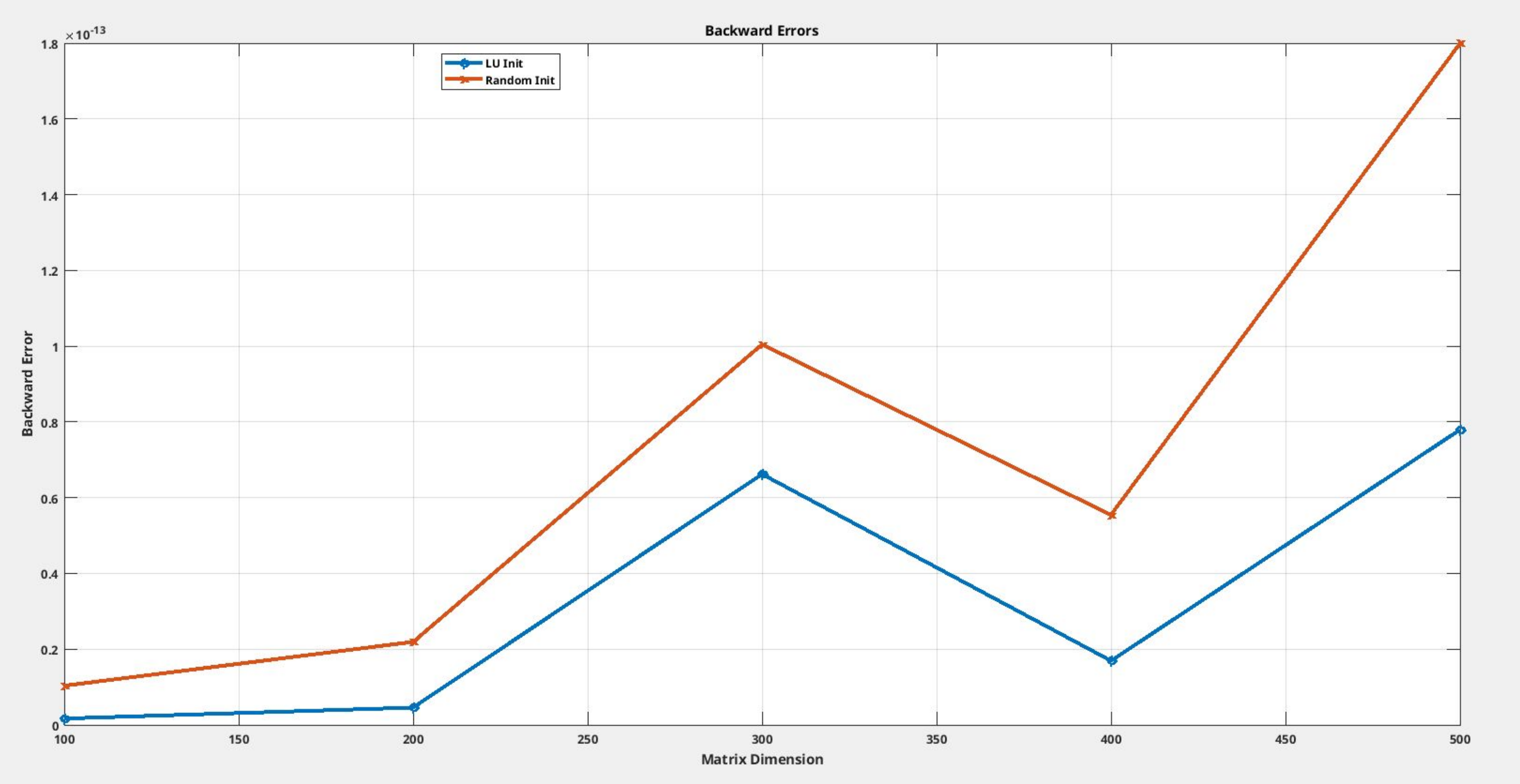
- Both approaches achieve **similar accuracy** after the same number of iterations.
- Backward and forward errors show **no major difference**.
- Getting FP16 LU initial solution is fast, but it is not free.

Forward Error



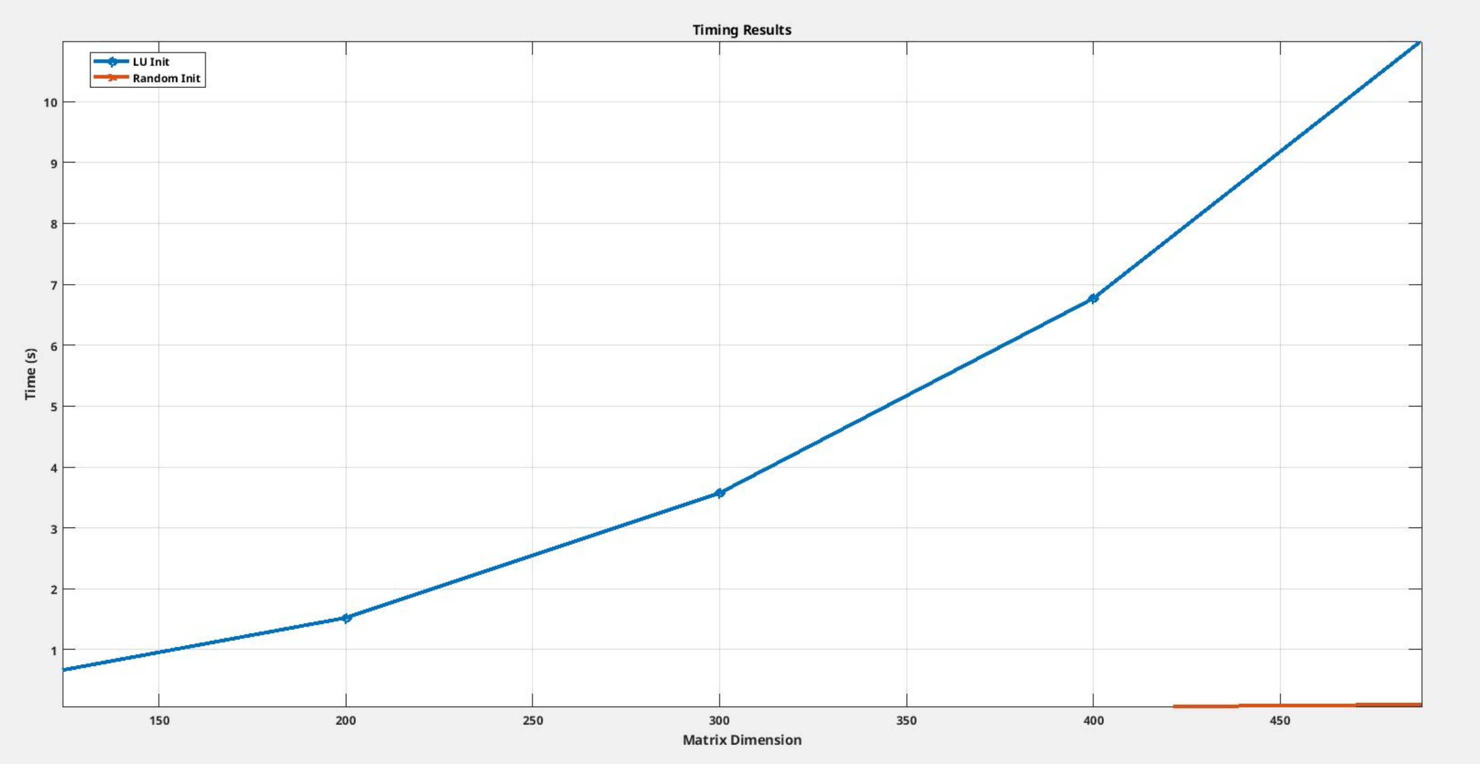
Simulated in Matlab

Backward Error



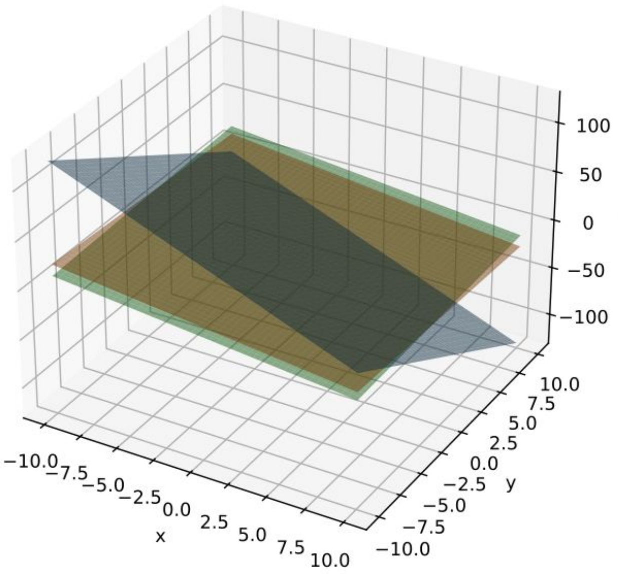
Simulated in Matlab

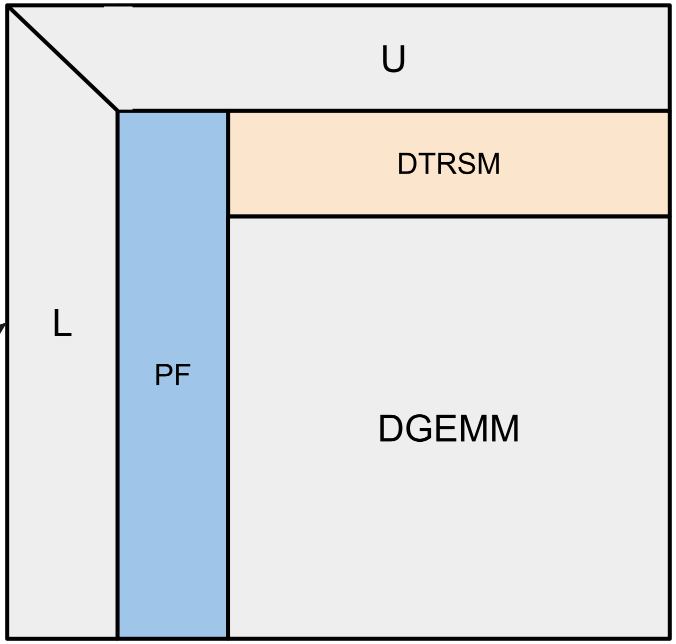
Execution Time



Simulated in Matlab

Block Implementation of LU Decomposition: Key Operations




U

DTRSM

L

PF

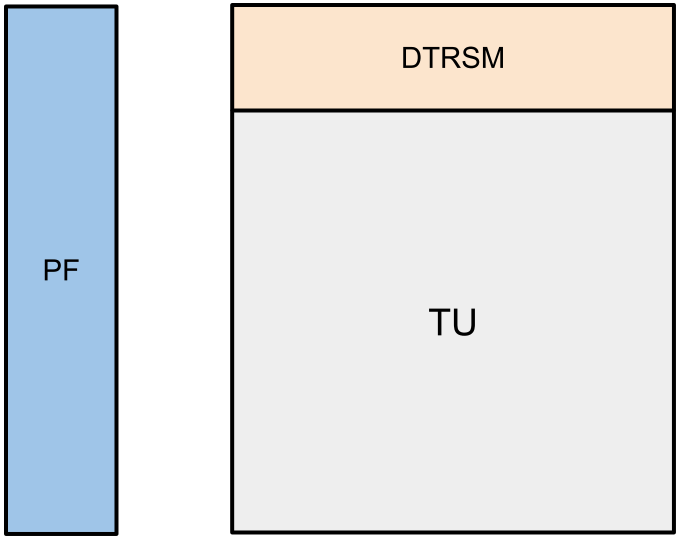
DGEMM

Geometrically as finding the intersection of hyperplanes.

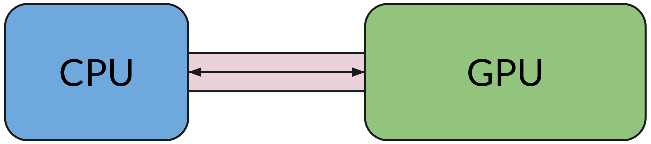
CPU and GPU in Sync for Blazing Fast Computation

Almost sequential

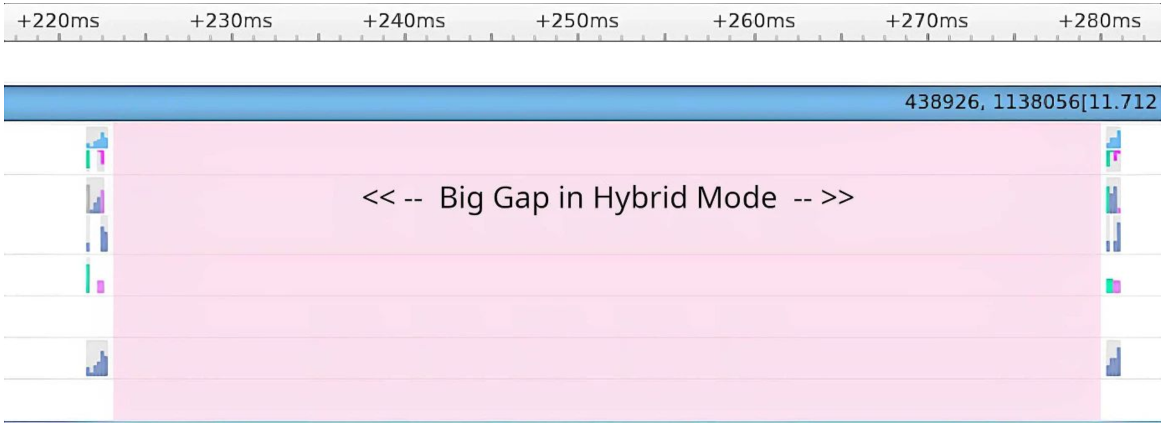
Highly parallel

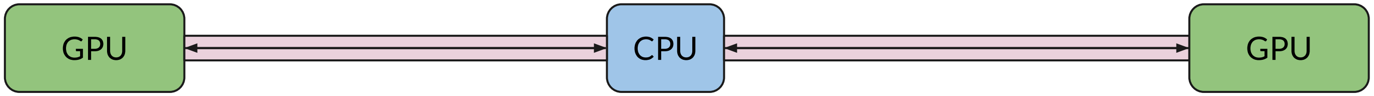


The CPU handles the sequential tasks, while the GPU accelerates parallel computations, leading to a faster overall process.



CPU and GPU in Sync for Blazing Fast Computation





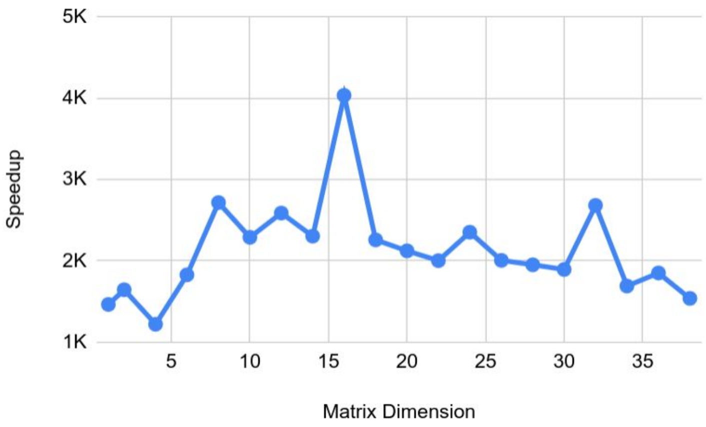
Observed a performance gap where the hybrid CPU-GPU setup is slower than expected. The integration of both resources results in a bottleneck [2].

Moving to Native GPU Mode: Improvement and GPU Underutilization

Shifted entirely to the GPU.

By offloading all tasks to the GPU, you can see a good performance improvement.

The GPU, optimized for parallel tasks, is now underutilized.



The "Two Lives" Concept

Imagine living your life twice: once in a **fast**, short version, and once in a **normal**, full version.

The **fast life** gives you a quick preview, but some details are missing. It makes you aware of the **consequences of your decisions**.

The **normal life** is more detailed but slower.

You can use your **short life memory** to reduce overthinking about the consequences of your decisions in normal life.

Compute it two times.

Can we make the computation even faster?

Do it two times.

Like the **fast life** doing computation in **reduced precision**, based on the outcome, **rearrange data** and start the **normal computation** without **overcomplicating**.

PRP and **MPF** Algorithms, repeat the computation 2 times.

Pre Pivoting (PRP):

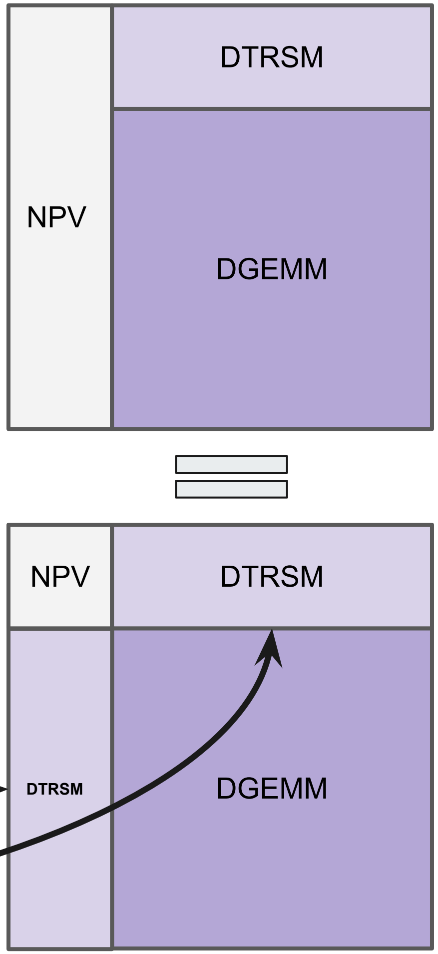
In PRP, we first perform the LU in FP16 (the ‘fast life’) to find the pivot list.

No search for pivots in FP64 columns.

Apply PF only to the **top square** section of the panel, and use a matrix **triangular solve** for the **remaining section** to take advantage of parallel GEMM.

Store the micro panel in shared memory and registers

Possibility to run two TRSMs in parallel.



Different Results with Panel vs. Whole Matrix:

For the Whole Matrix: When working with the whole matrix, we have two versions **[2]**:

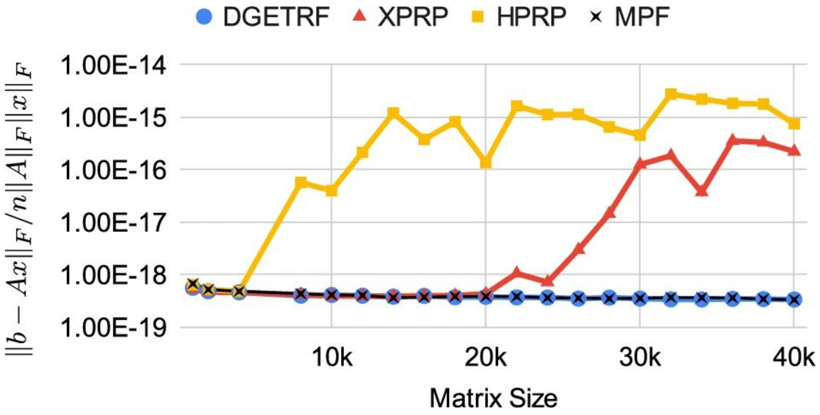
XPRP (using FP16/FP32) ;

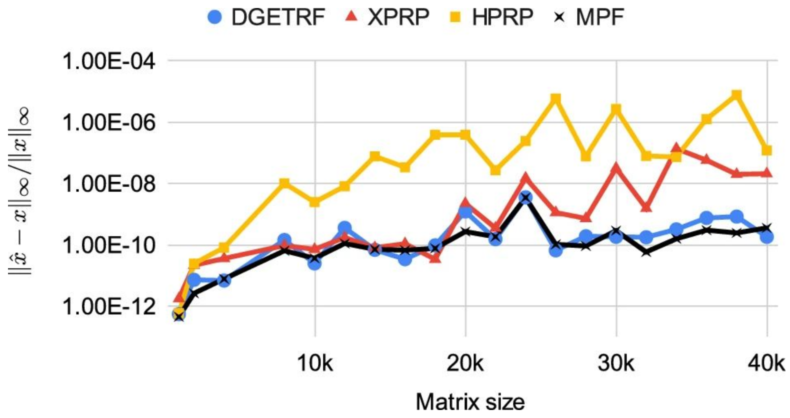
HPRP (using FP16 for pre-calculating LU over the entire matrix).

If we apply this approach to **just the next panel** of the matrix, the results are different. This approach is called **MPF**, and it is entirely in FP16 [2].

Offering different trade-offs in terms of speed and accuracy.

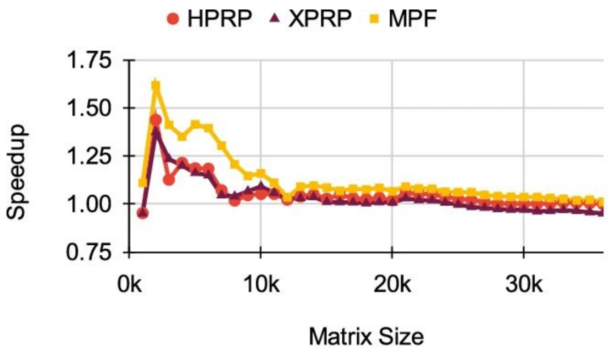
Results: Accuracy



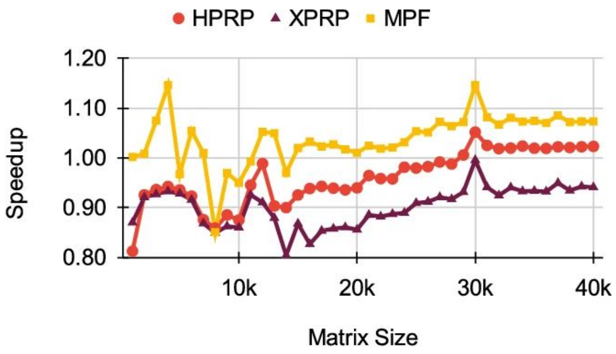


Data from [2]

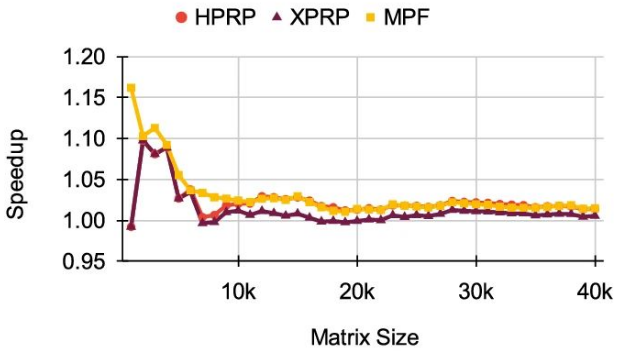
Results: Speedup



(a)



(b)



(c)

Data from [2]

References

1) **Haidar, A., Abdelfattah, A., Tomov, S. & Dongarra, J.** Harnessing GPU's Tensor Cores fast FP16 arithmetic to speed up mixed-precision iterative refinement solvers and achieve 74 Gflops/Watt on Nvidia V100. *GPU Technology Conference (GTC), Poster*, San Jose, CA (2018).
2) **Sahraneshinsamani, N., Catalán, S. & Herrero, J.R.** Mixed-precision pre-pivoting strategy for the LU factorization. *J Supercomput* 81, 87 (2025). https://doi.org/10.1007/s11227-024-06523-w

Harnessing Reduced Precision for Accurate and Efficient Scientific Computing in HPC

Fosdem 2025
2nd Feb, 2025

Nima Sahraneshin Samani

Supervisors
Sandra Catalan
José R. Herrero
José Ignacio Aliaga

UNIVERSITAT
JAUME·I

End!