Solving Sparse Linear Systems with Adaptive Precision GMRES

RAIM Meeting 2025

Emmanuel Agullo, Luc Giraud, Pierre Jolivet, Théo Mary & **Alexandre Tabouret**

LIP6 - Sorbonne University

November 6, 2025

Solving Linear System With GMRES

Let's consider the linear system $Ax = b$. GMRES generates a sequence of approximate solutions $(x_k)_{k \geq 1}$ from an initial solution x_0 that converges to the exact solution x^*:

Algorithm: GMRES(A, M^{-1}, b, x_0, ϵ_g)

$r_0 = b - Ax_0$;
$\beta = \|r_0\|$, $v_1 = r_0/\beta$, $k = 1$;
repeat
  $z_k = M^{-1}v_k$;
  $w_k = Az_k$;
  for $i = 1, \ldots, k$ **do**
    $h_{i,k} = v_i^T w_k$;
    $w_k = w_k - h_{i,k}v_i$;
  $h_{k+1,k} = \|w_k\|$;
  $v_{k+1} = w_k/h_{k+1,k}$;
  $V_k = [v_1, \ldots, v_k]$;
  $H_k = \{h_{i,j}\}_{1 \leq i,j \leq k+1}$;
  $y_k = \arg\min_y \|\beta e_1 - H_k y\|$;
  $k = k + 1$;
until $\eta_{A,b} \leq \epsilon_g$;
$d_k = V_k y_k$;
$x_k = x_0 + M^{-1}d_k$;

Solving Linear System With GMRES

Let's consider the linear system $Ax = b$.
GMRES generates a sequence of approximate solutions $(x_k)_{k\geq 1}$ from an initial solution x_0 that converges to the exact solution x^*:

- Build an orthogonal basis V_k of the Krylov subspace $span\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$;

Algorithm: GMRES(A, M^{-1}, b, x_0, ϵ_g)

$r_0 = b - Ax_0$;
$\beta = \|r_0\|$, $v_1 = r_0/\beta$, $k = 1$;
repeat
  $z_k = M^{-1}v_k$;
  $w_k = Az_k$;
  for $i = 1, \ldots, k$ **do**
    $h_{i,k} = v_i^T w_k$;
    $w_k = w_k - h_{i,k}v_i$;
  $h_{k+1,k} = \|w_k\|$;
  $v_{k+1} = w_k/h_{k+1,k}$;
  $V_k = [v_1, \ldots, v_k]$;
  $H_k = \{h_{i,j}\}_{1\leq i,j\leq k+1}$;
  $y_k = \arg\min_y \|\beta e_1 - H_k y\|$;
  $k = k + 1$;
until $\eta_{A,b} \leq \epsilon_g$;
$d_k = V_k y_k$;
$x_k = x_0 + M^{-1}d_k$;

Solving Linear System With GMRES

Let's consider the linear system $Ax = b$. GMRES generates a sequence of approximate solutions $(x_k)_{k \geq 1}$ from an initial solution x_0 that converges to the exact solution x^*:

- Build an orthogonal basis V_k of the Krylov subspace $span\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$;
- Chose x_k as the vector minimizing $\|r_k\| = \|b - Ax_k\|$;

Algorithm: GMRES(A, M^{-1}, b, x_0, ϵ_g)

$r_0 = b - Ax_0$;
$\beta = \|r_0\|$, $v_1 = r_0/\beta$, $k = 1$;
repeat
  $z_k = M^{-1} v_k$;
  $w_k = Az_k$;
  for $i = 1, \ldots, k$ **do**
    $h_{i,k} = v_i^T w_k$;
    $w_k = w_k - h_{i,k} v_i$;
  $h_{k+1,k} = \|w_k\|$;
  $v_{k+1} = w_k / h_{k+1,k}$;
  $V_k = [v_1, \ldots, v_k]$;
  $H_k = \{h_{i,j}\}_{1 \leq i,j \leq k+1}$;
  $y_k = \arg\min_y \|\beta e_1 - H_k y\|$;
  $k = k + 1$;
until $\eta_{A,b} \leq \epsilon_g$;
$d_k = V_k y_k$;
$x_k = x_0 + M^{-1} d_k$;

Solving Linear System With GMRES

Let's consider the linear system $Ax = b$. GMRES generates a sequence of approximate solutions $(x_k)_{k\geq 1}$ from an initial solution x_0 that converges to the exact solution x^*:

- Build an orthogonal basis V_k of the Krylov subspace $span\{r_0, Ar_0, \dots, A^{k-1}r_0\}$;
- Chose x_k as the vector minimizing $\|r_k\| = \|b - Ax_k\|$;
- Reiterate until the stopping criterion is satisfied. Here: $\eta_{a,b} = \frac{\|r_k\|}{\|A\|\|x_k\|+\|b\|}$

Algorithm: GMRES(A, M^{-1}, b, x_0, ϵ_g)

$r_0 = b - Ax_0$;
$\beta = \|r_0\|$, $v_1 = r_0/\beta$, $k = 1$;
repeat
  $z_k = M^{-1}v_k$;
  $w_k = Az_k$;
  for $i = 1, \dots, k$ **do**
    $h_{i,k} = v_i^T w_k$;
    $w_k = w_k - h_{i,k}v_i$;
  $h_{k+1,k} = \|w_k\|$;
  $v_{k+1} = w_k/h_{k+1,k}$;
  $V_k = [v_1, \dots, v_k]$;
  $H_k = \{h_{i,j}\}_{1\leq i,j\leq k+1}$;
  $y_k = \arg\min_y \|\beta e_1 - H_k y\|$;
  $k = k + 1$;
until $\eta_{A,b} \leq \epsilon_g$;
$d_k = V_k y_k$;
$x_k = x_0 + M^{-1}d_k$;

Mixed Precision GMRES

Usual approach to mixed precision GMRES: perform each core part of GMRES in a given precision.

- u_m: apply the preconditioner;
- u_a: apply the operator;
- u_g: rest of GMRES.

Algorithm: GMRES(A, M^{-1}, b, x_0, ϵ_g)

$r_0 = b - Ax_0$;
$\beta = \|r_0\|$, $v_1 = r_0/\beta$, $k = 1$;
repeat
  $z_k = M^{-1}v_k$;
  $w_k = Az_k$;
  for $i = 1, \ldots, k$ **do**
    $h_{i,k} = v_i^T w_k$;
    $w_k = w_k - h_{i,k}v_i$;
  $h_{k+1,k} = \|w_k\|$;
  $v_{k+1} = w_k/h_{k+1,k}$;
  $V_k = [v_1, \ldots, v_k]$;
  $H_k = \{h_{i,j}\}_{1 \leq i,j \leq k+1}$;
  $y_k = \arg\min_y \|\beta e_1 - H_k y\|$;
  $k = k + 1$;
until $\eta_{A,b} \leq \epsilon_g$;
$d_k = V_k y_k$;
$x_k = x_0 + M^{-1}d_k$;

Mixed Precision GMRES

Usual approach to mixed precision GMRES: perform each core part of GMRES in a given precision.

- u_m: apply the preconditioner;
- u_a: apply the operator;
- u_g: rest of GMRES.

Several papers offer different combinations of precisions, as well as a framework specifying the behavior of GMRES depending on the chosen precisions in [1].

[1] Alfredo Buttari et al. "Mixed precision strategies for preconditioned GMRES: a comprehensive analysis". working paper or preprint. May 2025. URL: https://hal.science/hal-05071696

Algorithm: GMRES(A, M^{-1}, b, x_0, ϵ_g)

$r_0 = b - Ax_0$;
$\beta = \|r_0\|$, $v_1 = r_0/\beta$, $k = 1$;
repeat
  $z_k = M^{-1}v_k$;
  $w_k = Az_k$;
  for $i = 1, \ldots, k$ **do**
    $h_{i,k} = v_i^T w_k$;
    $w_k = w_k - h_{i,k}v_i$;
  $h_{k+1,k} = \|w_k\|$;
  $v_{k+1} = w_k/h_{k+1,k}$;
  $V_k = [v_1, \ldots, v_k]$;
  $H_k = \{h_{i,j}\}_{1 \le i,j \le k+1}$;
  $y_k = \arg\min_y \|\beta e_1 - H_k y\|$;
  $k = k + 1$;
until $\eta_{A,b} \le \epsilon_g$;
$d_k = V_k y_k$;
$x_k = x_0 + M^{-1}d_k$;

Our goal and approach

The base idea was to improve GMRES performance by introducing a mixed precision Sparse Matrix-Vector product (SpMV), while ensuring convergence.

That requires to:

1. dispose of a mixed precision SpMV algorithm;
2. find a way to ensure convergence despite the use of low precisions.

Adapt SpMV

Adapt SpMV [2]:

- Algorithm designed by the Team PEQUAN from LIP6;
- Perform the SpMV $A * v$ in mixed precision;

[2] Stef Graillat et al. "Adaptive Precision Sparse Matrix–Vector Product and Its Application to Krylov Solvers". In: *SIAM Journal on Scientific Computing* 46.1 (2024), pp. C30–C56. DOI: 10.1137/22M1522619

Adapt SpMV

`Adapt SpMV` [2]:

- Algorithm designed by the Team PEQUAN from LIP6;
- Perform the SpMV $A * v$ in mixed precision;
- Each element of A is **represented with a precision proportional to its magnitude**;

[2] Stef Graillat et al. "Adaptive Precision Sparse Matrix–Vector Product and Its Application to Krylov Solvers". In: *SIAM Journal on Scientific Computing* 46.1 (2024), pp. C30–C56. DOI: 10.1137/22M1522619

Adapt SpMV

`Adapt SpMV` [2]:

- Algorithm designed by the Team PEQUAN from LIP6;
- Perform the SpMV $A * v$ in mixed precision;
- Each element of A is **represented with a precision proportional to its magnitude**;
- Uses **custom precision formats**: BF16, FP24, FP32, FP40, FP48, FP56, FP64 and FP00;

[2] Stef Graillat et al. "Adaptive Precision Sparse Matrix–Vector Product and Its Application to Krylov Solvers". In: *SIAM Journal on Scientific Computing* 46.1 (2024), pp. C30–C56. DOI: 10.1137/22M1522619

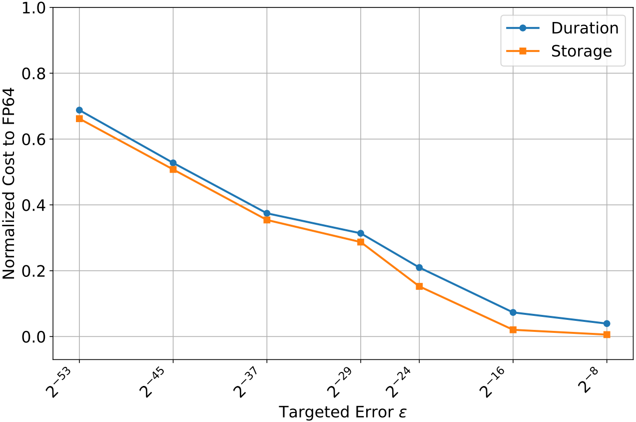
Adapt SpMV

Adapt SpMV [2]:

- Algorithm designed by the Team PEQUAN from LIP6;
- Perform the SpMV $A * v$ in mixed precision;
- Each element of A is **represented with a precision proportional to its magnitude**;
- Uses **custom precision formats**: BF16, FP24, FP32, FP40, FP48, FP56, FP64 and FP00;
- The parameter ϵ specifies to the algorithm the desired relative error on the mixed precision representation of A.

[2] Stef Graillat et al. "Adaptive Precision Sparse Matrix–Vector Product and Its Application to Krylov Solvers". In: *SIAM Journal on Scientific Computing* 46.1 (2024), pp. C30–C56. DOI: 10.1137/22M1522619

Adapt SpMV: Performance on the matrix Long_Coup_dt6



→ Execution duration is proportional to storage.

How To Ensure Convergence?

Mixed precision computations deteriorate the quality of the results. That can be seen as perturbations, which are critical in many applications.

For GMRES, perturbations on the SpMV are represented as a perturbation matrix E_k such that the computation performed is no longer $A * v$ but $(A + E_k) * v$.

How can we set ϵ to prevent any negative impact on GMRES behavior. One might want to simply set ϵ equal to the targeted tolerance for the GMRES solution However, we can do better.

Convergence in Backward Error of Relaxed GMRES [3]

Theorem (Convergence of Relaxed GMRES)

Let x_0 (initial solution) and $c = 0.5$ a constant. Let ϵ_g, the target tolerance for GMRES solution be any positive real number. Then supposed for all k

$$\frac{||E_k||}{||A||} \leq \frac{1}{n * \kappa(A)} \min\left(c, (1-c)\frac{||b||}{||r_{k-1}||}\epsilon_g\right),$$

Then the convergence according to the stopping criterion $\eta_{A,b}$ is ensured.

[3] Luc Giraud, Serge Gratton, and Julien Langou. "Convergence in backward error of relaxed GMRES". In: *SIAM J. Scientific Computing* 29 (Jan. 2007), pp. 710–728. DOI: 10.1137/040608416

Convergence in Backward Error of Relaxed GMRES [3]

Theorem (Convergence of Relaxed GMRES)

Let x_0 (initial solution) and $c = 0.5$ a constant. Let ϵ_g, the target tolerance for GMRES solution be any positive real number. Then supposed for all k

$$\frac{\|E_k\|}{\|A\|} \leq \frac{1}{n * \kappa(A)} \min\left(c, (1-c)\frac{\|b\|}{\|r_{k-1}\|}\epsilon_g\right),$$

Then the convergence according to the stopping criterion $\eta_{A,b}$ is ensured.

$\rightarrow$ $\|E_k\|$ is inversely proportional to $\|r_{k-1}\|$, meaning that the error is permitted to increase in proportion to the reduction of the residual.

[3] Luc Giraud, Serge Gratton, and Julien Langou. "Convergence in backward error of relaxed GMRES". In: *SIAM J. Scientific Computing* 29 (Jan. 2007), pp. 710–728. DOI: 10.1137/040608416

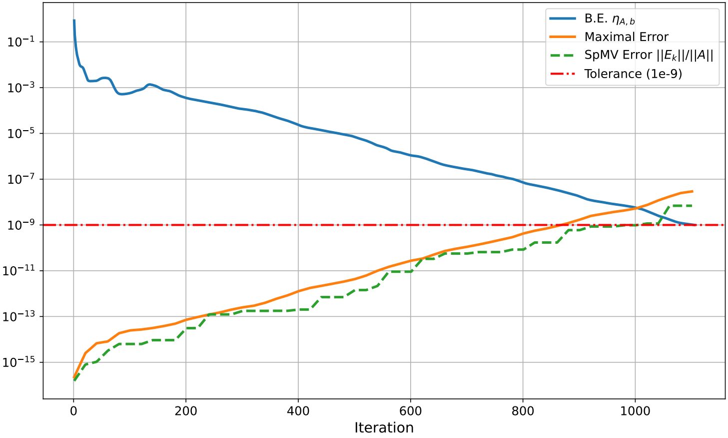
Mixed and Adaptive Precision GMRES

- At the start of each iteration, adapt the precision of A;
- Perform the SpMV in mixed precision.

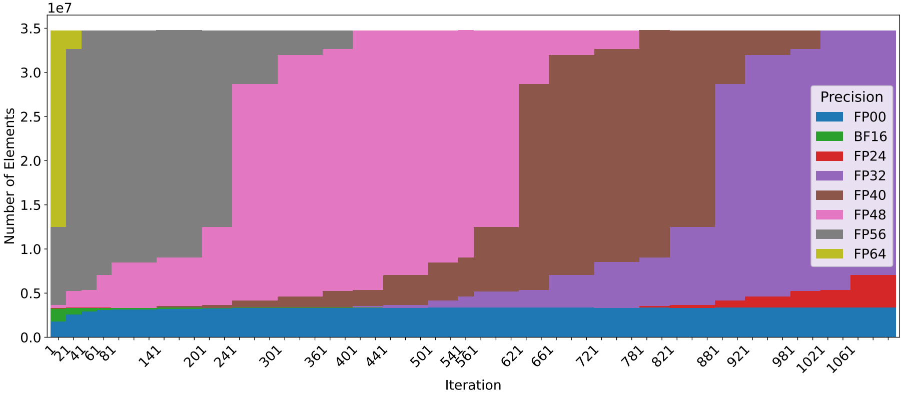
Algorithm: GMRES(A, M^{-1}, b, x_0, ϵ_g)

$r_0 = b - Ax_0$;
$\beta = \|r_0\|$, $v_1 = r_0/\beta$, $k = 1$;
repeat
  ϵ = Compute target precision from r_k;
  Adjust the matrix A to precision ϵ;
  $z_k = M^{-1} v_k$;
  w_k = AdaptSpMV(A, z_k);
  for $i = 1, \ldots, k$ **do**
    $h_{i,k} = v_i^T w_k$;
    $w_k = w_k - h_{i,k} v_i$;
  $h_{k+1,k} = \|w_k\|$;
  $v_{k+1} = w_k / h_{k+1,k}$;
  $V_k = [v_1, \ldots, v_k]$;
  $H_k = \{h_{i,j}\}_{1 \le i,j \le k+1}$;
  $y_k = \arg\min_y \|\beta e_1 - H_k y\|$;
  $k = k + 1$;
until $\eta_{A,b} \le \epsilon_g$;
$d_k = V_k y_k$;
$x_k = x_0 + M^{-1} d_k$;

Convergence: Matrix ss



Composition of the Matrix ss



Experiment Set Up: Restarted GMRES

Main limitation of GMRES: Cost of orthogonalizing the Krylov basis $\mathcal{O}(nk)$.

For that reason, we used Restarted GMRES or GMRES(m):

- Run GMRES for m iterations $\rightarrow$ approximate solution;
- Restart GMRES with that solution as the new initial guess;
- Repeat until convergence.

Experiment Set Up: Restarted GMRES

Main limitation of GMRES: Cost of orthogonalizing the Krylov basis $\mathcal{O}(nk)$.

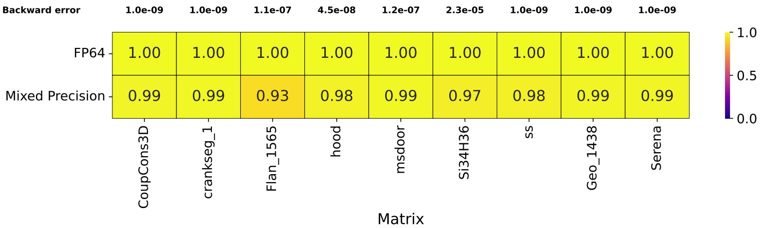
For that reason, we used Restarted GMRES or GMRES(m):

- Run GMRES for m iterations $\rightarrow$ approximate solution;
- Restart GMRES with that solution as the new initial guess;
- Repeat until convergence.

$\rightarrow$ Limits Krylov subspace basis size: better memory.
$\rightarrow$ Prevents the orthogonalization from becoming too expensive: better performance.

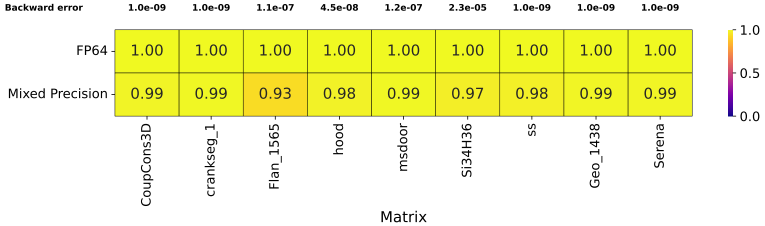
Performance

Duration of mixed and adaptive precision GMRES(80), normalized by the duration of the double precision version.



Performance

Duration of mixed and adaptive precision GMRES(80), normalized by the duration of the double precision version.



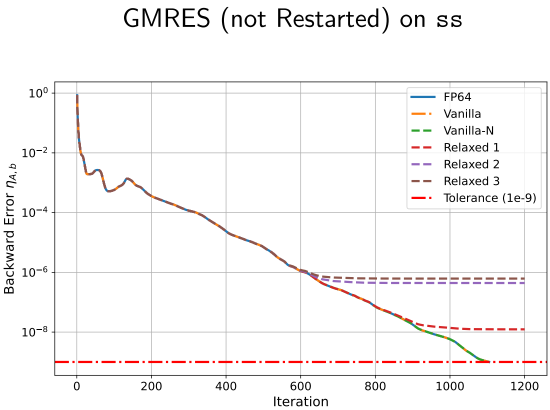
→ Around 1 to 3% faster than regular GMRES only: trade-off between the benefit from computing the mixed precision SpMV and the cost of regularly adapting the precision of A.

Faster SpMV: Heuristics

The theorem is very "safe" and in many cases, exceeding the boundary does not necessarily prevent convergence. Therefore multiple heuristics were designed:

- Conservative Heuristics
  - Vanilla: $\frac{\|E_k\|}{\|A\|} \leq \frac{1}{n*\kappa(A)} \min\left(c, (1-c)\frac{\|b\|}{\|r_{k-1}\|}\epsilon_g\right)$;
  - VanillaN: $\frac{\|E_k\|}{\|A\|} \leq \frac{1}{n} \min\left(c, (1-c)\frac{\|b\|}{\|r_{k-1}\|}\epsilon_g\right)$;
- Relaxed heuristics
  - Relaxed 1: $\frac{\|E_k\|}{\|A\|} \leq \max\left(\epsilon_g, c\frac{\|b\|}{\|r_{k-1}\|}\frac{\epsilon_g}{n}\right)$;
  - Relaxed 2: $\frac{\|E_k\|}{\|A\|} \leq \max\left(\epsilon_g, c\frac{\|b\|}{\|r_{k-1}\|}\frac{\epsilon_g}{\kappa(A)}\right)$;
  - Relaxed 3: $\frac{\|E_k\|}{\|A\|} \leq \max\left(\epsilon_g, c\frac{\|b\|}{\|r_{k-1}\|}\epsilon_g\right)$.

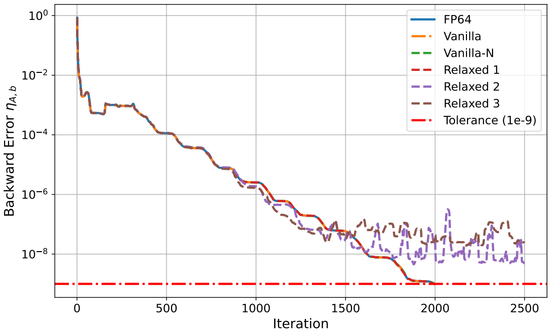
Convergence Problems



Relaxed heuristics end up stagnating because they are introducing too much perturbations and GMRES breaks.

Restarting to Succeed

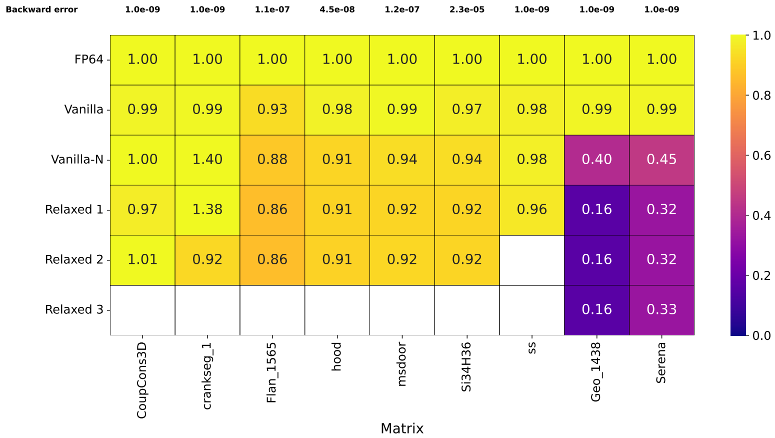
GMRES(80) on ss



With restarts, relaxed heuristics may converge. By restarting often enough, we are preventing perturbations from accumulating too much.

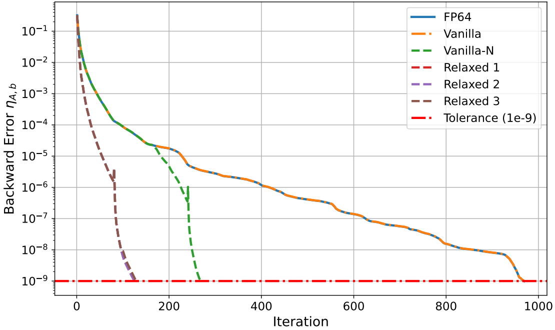
Results on a Few Matrices

Duration of GMRES(80) with the different heuristics, normalized by the duration of the double precision version.



Mixed Precision Induced Fast Convergence Phenomenon

GMRES(80) on `Serena`



Perturbations can positively impact GMRES.
They can potentially make the system easier to solve for GMRES.

Conclusion & Future Works

Mixed and adaptive precision GMRES:

- Dynamically adapts the precision of each coefficient of A;
- Improves performance using the `Adapt SpMV` algorithm;
- Manages to reduce GMRES duration by 5 to 15%.

Conclusion & Future Works

Mixed and adaptive precision GMRES:

- Dynamically adapts the precision of each coefficient of A;
- Improves performance using the `Adapt SpMV` algorithm;
- Manages to reduce GMRES duration by 5 to 15%.

Future (possible) works:

- Preliminary paper;
- Mixed precision preconditioners;
- What about other Krylov methods?

Conclusion & Future Works

Mixed and adaptive precision GMRES:

- Dynamically adapts the precision of each coefficient of A;
- Improves performance using the `Adapt SpMV` algorithm;
- Manages to reduce GMRES duration by 5 to 15%.

Future (possible) works:

- Preliminary paper;
- Mixed precision preconditioners;
- What about other Krylov methods?

Thank you for your attention. Any questions?

References

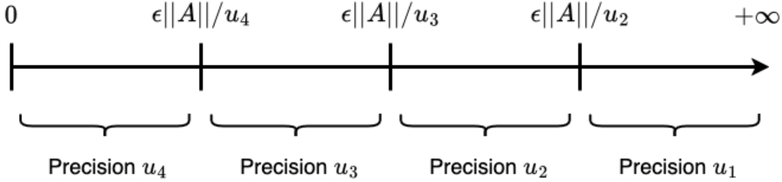
[1] Alfredo Buttari et al. "Mixed precision strategies for preconditioned GMRES: a comprehensive analysis". working paper or preprint. May 2025. URL: https://hal.science/hal-05071696.

[2] Stef Graillat et al. "Adaptive Precision Sparse Matrix–Vector Product and Its Application to Krylov Solvers". In: *SIAM Journal on Scientific Computing* 46.1 (2024), pp. C30–C56. DOI: 10.1137/22M1522619.

[3] Luc Giraud, Serge Gratton, and Julien Langou. "Convergence in backward error of relaxed GMRES". In: *SIAM J. Scientific Computing* 29 (Jan. 2007), pp. 710–728. DOI: 10.1137/040608416.

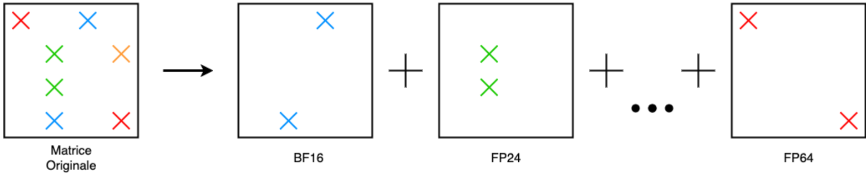
Adapt SpMv : Precisions

Elements of A are attributed to buckets of various precision, depending on their amplitude. Buckets bounds are computed using $\|A\|$, ϵ and the available precisions $u_1 < \cdots < u_k$.

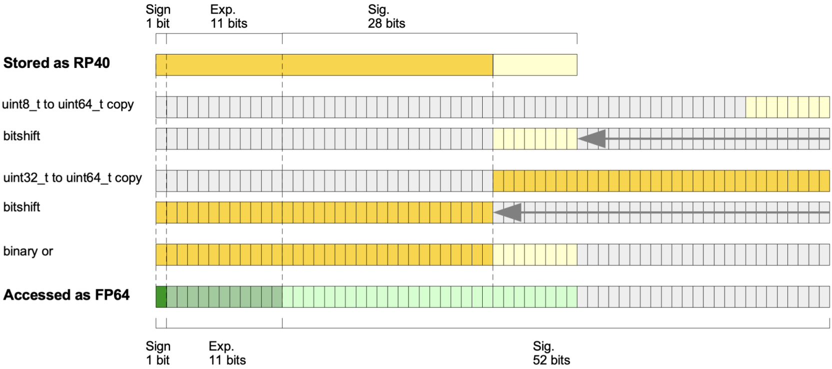


Adapt SpMv : Format & Storage

Each element is stored in a matrix corresponding to its precision. The different matrices are in the CSR (*Compressed Sparse Row*) format.



Adapt SpMv : Decompression



Experiment Set Up: Code & Environment

Code:

- Adapt SpMV: Multi-threaded (OpenMP);
- Composyx: High level linear algebra library, linked to multi-threaded Intel MKL (used for matrix and vector operations);
- Using Guix for reproducibility and deployment.

Environment:

- Bora nodes of PlaFRIM: 2x 18-core Intel CascadeLake & 192 GB of memory;
- Using only one socket (one CPU / 18 threads);
- Run using `numactl -interleave=all`.