Accelerating GMRES via Mixed Precision

Neil Lindquist, Piotr Luszczek, Jack Dongarra

12th JLESC Workshop

February 25th, 2020

INNOVATIVE COMPUTING LABORATORY

THE UNIVERSITY OF TENNESSEE KNOXVILLE

GMRES

- General purpose, sparse linear solver
  - Iterative, Krylov solver
- Memory bound performance
  - Mix single and double precision

GMRES Algorithm

$\text{GMRES}_{res}(\boldsymbol{A}, \boldsymbol{x_0}, \boldsymbol{b}, \boldsymbol{M}^{-1})$ — Computing $\boldsymbol{Ax} = \boldsymbol{b}$. $\boldsymbol{A}^{-1} \approx \boldsymbol{M}^{-1}$

for $k = 0, 1, 2, \ldots$ — Restarts

$\boldsymbol{r_k} \leftarrow \boldsymbol{b} - \boldsymbol{Ax_k}$

$\boldsymbol{z_k} \leftarrow \boldsymbol{M}^{-1}\boldsymbol{r_k}$

$\beta \leftarrow \|\boldsymbol{z_k}\|_2$

$\boldsymbol{V}_{:,0} \leftarrow \boldsymbol{z_k}/\beta$

$\boldsymbol{s} \leftarrow [\beta, 0, 0, \ldots, 0]^T$

for j $= 0, 1, 2, \ldots$ — Iteration count

$\boldsymbol{w} \leftarrow \boldsymbol{M}^{-1}\boldsymbol{AV}_{:,j}$

$\boldsymbol{w}, \boldsymbol{H}_{:,j} \leftarrow orthogonalize(\boldsymbol{w}, \boldsymbol{V}_{:,j})$

$\boldsymbol{H}_{j+1,j} \leftarrow \|\boldsymbol{w}\|_2$

$\boldsymbol{V}_{:,j+1} \leftarrow \boldsymbol{w}/\|\boldsymbol{w}\|_2$

$\boldsymbol{H}_{:,j} \leftarrow \boldsymbol{G_0 G_1} \ldots \boldsymbol{G_{j-1} H}_{:,j}$

$\boldsymbol{G_j} \leftarrow rotation\_matrix(\boldsymbol{H}_{:,j})$

$\boldsymbol{H}_{:,j} \leftarrow \boldsymbol{G_j H}_{:,j}$

$\boldsymbol{s} \leftarrow \boldsymbol{G_j s}$

$\boldsymbol{u_k} \leftarrow \boldsymbol{VH}^{-1}\boldsymbol{s}$

$\boldsymbol{x_{k+1}} \leftarrow \boldsymbol{x_k} + \boldsymbol{u_k}$

GMRES Algorithm

$\text{GMRES}_{res}(\boldsymbol{A}, \boldsymbol{x_0}, \boldsymbol{b}, \boldsymbol{M}^{-1})$

for $k = 0, 1, 2, \ldots$

Double: $\boldsymbol{r_k} \leftarrow \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x_k}$

Single:

$\boldsymbol{z_k} \leftarrow \boldsymbol{M}^{-1}\boldsymbol{r_k}$

$\beta \leftarrow \|\boldsymbol{z_k}\|_2$

$\boldsymbol{V}_{:,0} \leftarrow \boldsymbol{z_k}/\beta$

$\boldsymbol{s} \leftarrow [\beta, 0, 0, \ldots, 0]^T$

for j $= 0, 1, 2, \ldots$

$\quad \boldsymbol{w} \leftarrow \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{V}_{:,j}$

$\quad \boldsymbol{w}, \boldsymbol{H}_{:,j} \leftarrow orthogonalize(\boldsymbol{w}, \boldsymbol{V}_{:,j})$

$\quad \boldsymbol{H}_{j+1,j} \leftarrow \|\boldsymbol{w}\|_2$

$\quad \boldsymbol{V}_{:,j+1} \leftarrow \boldsymbol{w}/\|\boldsymbol{w}\|_2$

$\quad \boldsymbol{H}_{:,j} \leftarrow \boldsymbol{G_0}\boldsymbol{G_1} \ldots \boldsymbol{G_{j-1}}\boldsymbol{H}_{:,j}$

$\quad \boldsymbol{G_j} \leftarrow rotation\_matrix(\boldsymbol{H}_{:,j})$

$\quad \boldsymbol{H}_{:,j} \leftarrow \boldsymbol{G_j}\boldsymbol{H}_{:,j}$

$\quad \boldsymbol{s} \leftarrow \boldsymbol{G_j}\boldsymbol{s}$

$\boldsymbol{u_k} \leftarrow \boldsymbol{V}\boldsymbol{H}^{-1}\boldsymbol{s}$

Double: $\boldsymbol{x_{k+1}} \leftarrow \boldsymbol{x_k} + \boldsymbol{u_k}$

Computing $\boldsymbol{Ax} = \boldsymbol{b}$. $\boldsymbol{A}^{-1} \approx \boldsymbol{M}^{-1}$

Restarts

Iteration count

GMRES Simplified Algorithm

$$\text{GMRES}_{res}(\boldsymbol{A}, \boldsymbol{x_0}, \boldsymbol{b}, \boldsymbol{M}^{-1})$$

$$\text{for } \ k = 0, 1, 2, \ldots$$

Double: $\boldsymbol{r_k} \leftarrow \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x_k}$

Single: $\boldsymbol{u_k} \leftarrow \text{GMRES}_{no\ res}(\boldsymbol{A}, \vec{\boldsymbol{0}}, \boldsymbol{r_k}, \boldsymbol{M}^{-1})$

Double: $\boldsymbol{x_{k+1}} \leftarrow \boldsymbol{x_k} + \boldsymbol{u_k}$

GMRES Simplified Algorithm

$\text{GMRES}_{res}(\boldsymbol{A}, \boldsymbol{x_0}, \boldsymbol{b}, \boldsymbol{M}^{-1})$

for $k = 0, 1, 2, \ldots$

Double: $\boldsymbol{r_k} \leftarrow \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x_k}$

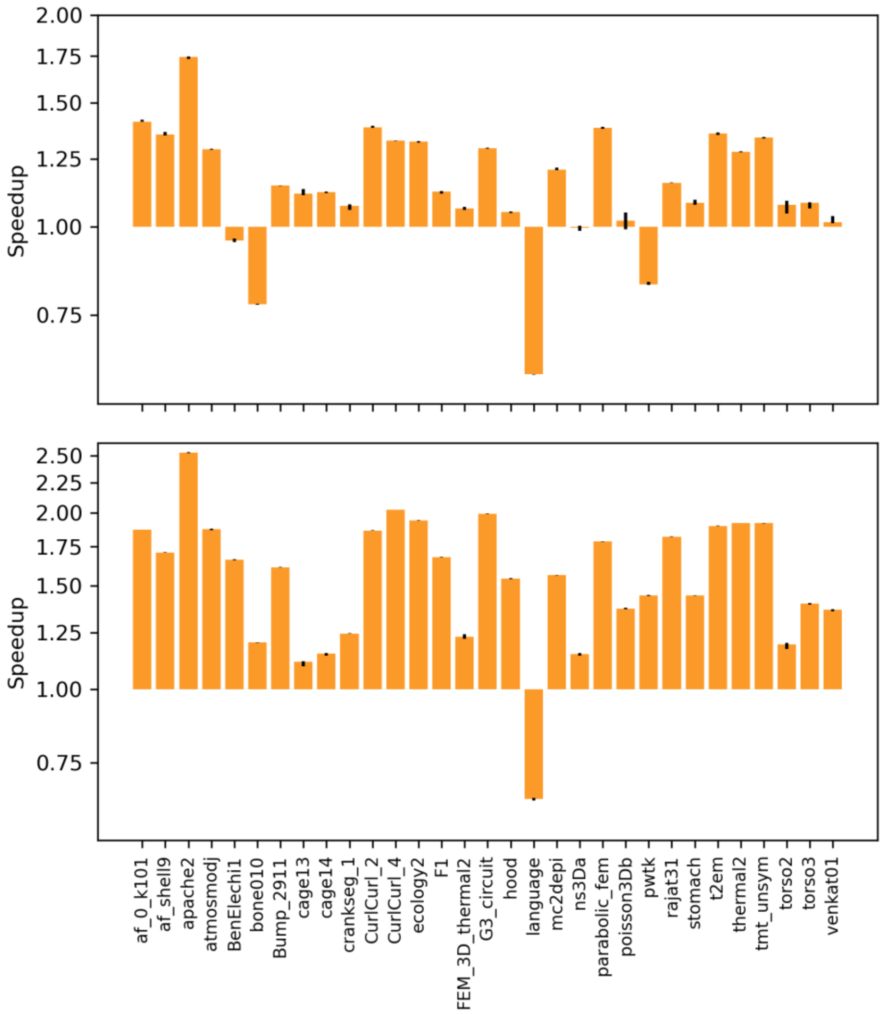
Single: $\boldsymbol{u_k} \leftarrow \boldsymbol{A}^{-1}\, \boldsymbol{r_k}$

Double: $\boldsymbol{x_{k+1}} \leftarrow \boldsymbol{x_k} + \boldsymbol{u_k}$

Performance

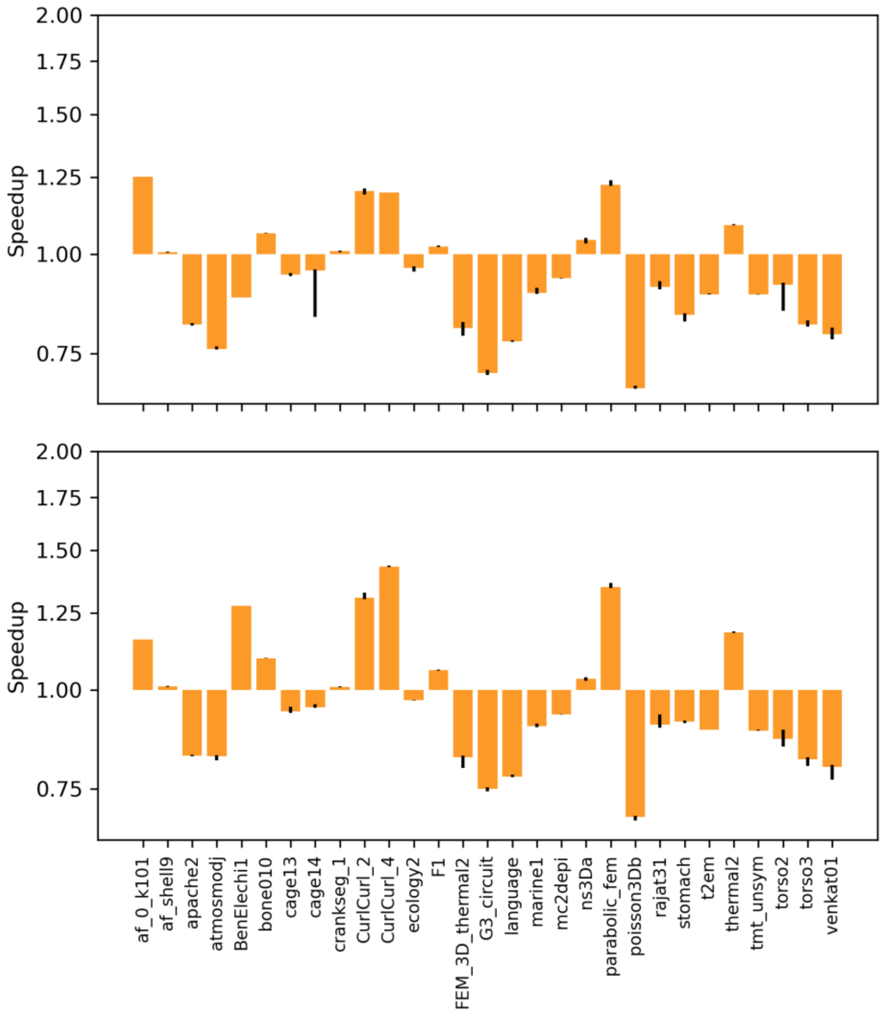
- Target accuracy $10^{-10} = \frac{\|b-Ax\|_2}{\|A\|_F\|x\|_2+\|b\|_2}$
- Restart strategies:
  I. 100 inner iterations
  II. 100 inner iterations or residual estimate of 10^{-10}
  III. First: 100 inner iterations or residual estimate of 10^{-6}
  Then: same number of inner iterations
- 20-core Haswell node with NVIDIA V100 GPU
  - cuSparse, cuBLAS, Kokkos
- CSR matrix format

Performance – Scalar Jacobi



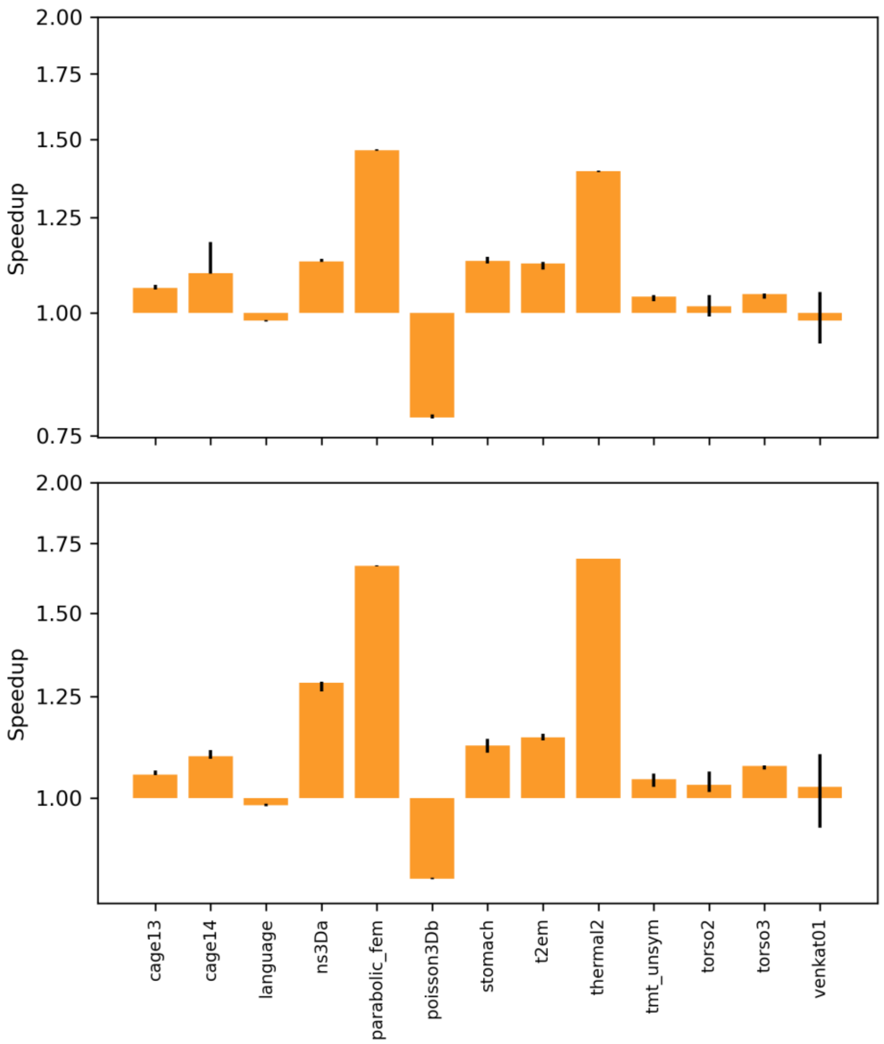
- Speedups
  - Median time of 3 run
  - 3 runs
  - Error bars: mins and maxes
- Geometric mean of speedup
  - MGS: 14%
  - CGSR: 54%

Performance – ILU(0)



- Speedups
  - Median time of 3 run
  - 3 runs
  - Error bars: mins and maxes
- Geometric mean of speedup
  - MGS: -7%
  - CGSR: -4%

Performance – ILU(0) with Jacobi Solves



- ILU(0) w/ 5 Jacobi iterations for each triangular solve
- Speedups
  - Median time of 3 run
  - 3 runs
  - Error bars: mins and maxes
- Geometric mean of speedup
  - MGS: 8%
  - CGSR: 14%

Future Directions

- Choice of low-precision
  - Half, Bfloat16
  - Compression
- Distributed systems
- Other Krylov methods
- Applications

Conclusions

- When restarted, mixed-precision GMRES often outperforms double-precision GMRES

Extra Slides

Test Configuration Details

- CUDA 10.2.199, Kokkos 3.1.01, GCC 7.3.0
- https://bitbucket.org/icl/mixed-precision-gmres
  - tag TPDS-perf

Publications

- N. Lindquist, P. Luszczek, and J. Dongarra, “Improving the performance of the GMRES method using mixed-precision techniques,” in Driving Scientific and Engineering Discoveries through the Convergence of HPC, Big Data and AI. DOI: 10.1007/978-3-030-63393-6_4
- [Submitted] N. Lindquist, P. Luszczek, and J. Dongarra, “Accelerating restarted GMRES with mixed precision arithmetic,” in Transactions on Parallel and Distributed Systems.

Effect on Convergence: Configuration

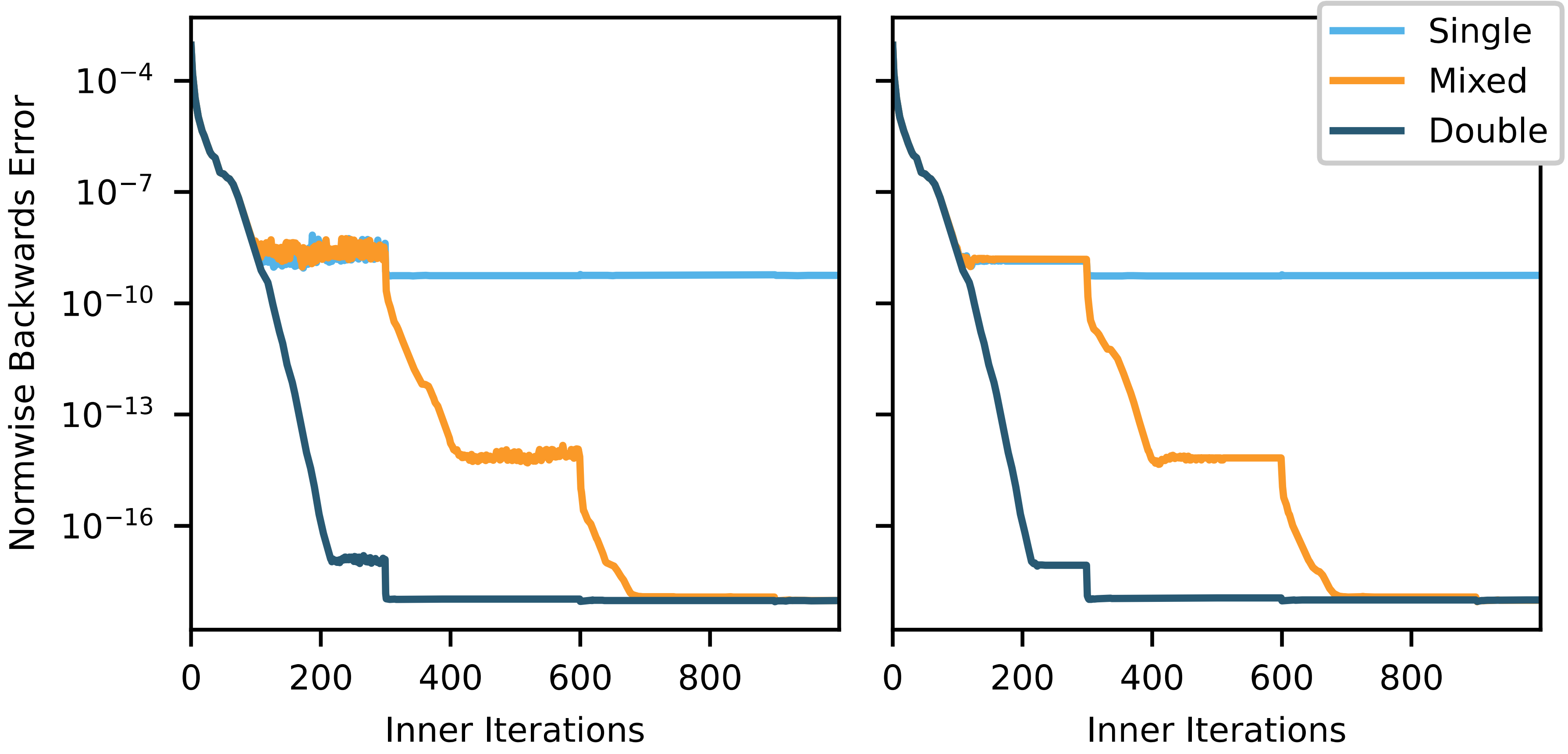
- ILU(0) preconditioner (M^{-1})
- CSR matrix format
- Custom, mixed precision kernels w/ Kokkos
- 20-core Haswell node
  - 2x Intel® Xeon® E5-2650 v3 processors

Effect on Convergence: Configuration

- airfoil_2d from SuiteSparse collection
  - $n = 14{,}214$
  - $nnz = 259{,}688$
  - $\kappa_2 = 1.8 \cdot 10^6$
- Error if GMRES stopped

$$\frac{\|b - Ax\|_2}{\|A\|_F \|x\|_2 + \|b\|_2}$$

Accuracy results



Modified Gram-Schmidt Orthogonalization (MGS)

Classical Gram-Schmidt with Reorthogonalization (CGSR)