SAND2022-8038D

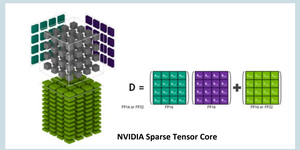
*Sandia National Laboratories*

# Mixed Precision Strategies for GMRES in Trilinos

Jennifer Loe, Christian Glusa, Ichitaro Yamazaki, Erik Boman, Sivasankaran Rajamanickam

## Motivation:

- Linear solve Ax=b can be the most expensive kernel in application simulations.
- Linear solvers are communication-bound (bandwidth-bound) algorithms.
- Reduce the costs of data movement and calculation by using lower-precision data storage and arithmetic.
- Leverage new specialized (ML/AI) hardware that favors low precision computations.
- How to use low precision calculations while maintaining high-precision accuracy needed by applications?




## Technical Approach:

**Algorithm: Mixed Precision GMRES with Iterative Refinement (GMRES-IR) for solving *Ax=b*:**

Compute residual *b-Ax*. (64-bit precision)

While not converged:

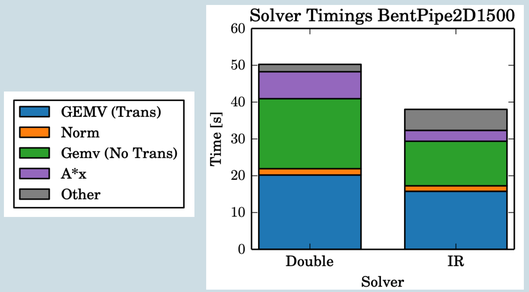
- Run GMRES linear solver 50 steps. (32-bit precision)
- Compute *b-Ax* to seed next GMRES run. (64-bit)

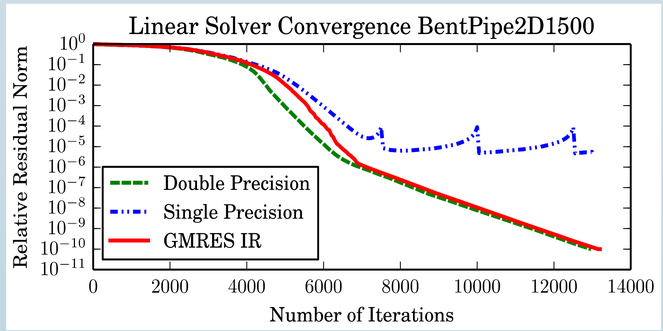
- Known algorithm for managing roundoff error [1,2,3]. We tested performance and reliability on NVIDIA V100 GPU.
- Implemented algorithm in Trilinos solvers package with a Kokkos backend for GPU performance.


- *[1] Improving the performance of the GMRES method using mixed-precision techniques.* Neil Lindquist, Piotr Luszczek, and Jack Dongarra.
- *[2] Mixed precision iterative refinement methods for linear systems: Convergence analysis based on Krylov subspace methods.* Hartwig Anzt, Vincent Heuveline, and Bjorn Rocker.
- *[3] Accelerating the solution of linear systems by iterative refinement in three precisions.* Erin Carson and Nicholas J. Higham.


## Results Summary:

- Improvements in speed of orthogonalization and sparse matrix-vector product kernels give total speedup of about 30% over GMRES (64-bit). Tested on V100 GPUs.
- Convergence of GMRES-IR (mixed) follows GMRES (64-bit) for both preconditioned and non-preconditioned linear systems!
- Modeled 2.5x sparse matrix-vector product speedup due to improved L2 cache use.




## Impacts & Successes to Date:

- Part of ECP xSDK multiprecision effort: Joint effort to propagate mixed precision software and algorithms across many labs. [4]
- Workshop paper published on GMRES-IR GPU performance. [5]
- Designed GMRES-IR benchmark code to test capabilities of new HPC systems for mixed precision linear solvers.
- Half precision support available in Kokkos and Kokkos Kernels.
- Mixed precision preconditioning available in Trilinos.
- Upcoming: Trilinos GMRES-IR solver software available to applications (Kokkos + MPI).


*[4] A Survey of Numerical Linear Algebra Methods Utilizing Mixed Precision Arithmetic, A. Abdelfattah, et. al. The International Journal of High Performance Computing Applications, March 2021.*
*[5] Experimental Evaluation of Multiprecision Strategies for GMRES on GPUs, J. Loe, C. Glusa, I. Yamazaki, S. Rajamanickam, and E. Boman, 2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 469-478, 2021.*



Funding: Exascale Computing Project (ECP): xSDK Multiprecision Focus Effort
[Sandia Portion: 3-year effort at ~0.55 FTEs per year] -- CIS Mathematics, Algorithms, and Simulations Area (MAS) --