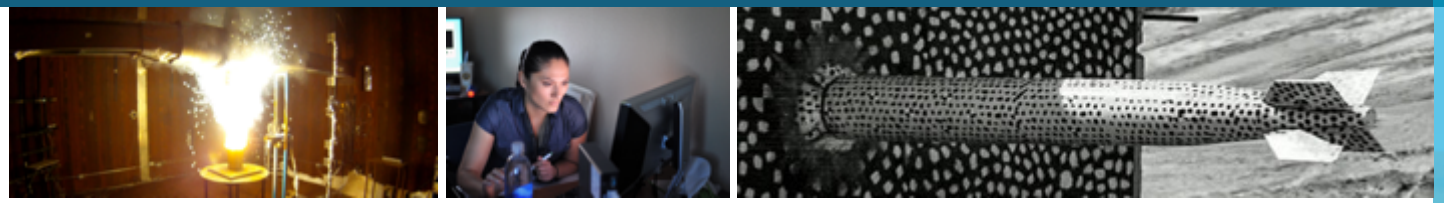
SAND2020-13230C

Sandia National Laboratories

Mixed-Precision GMRES in Trilinos



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

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2020-

ECP EXASCALE COMPUTING PROJECT

Mixed-Precision Krylov Solvers in Trilinos:

- GMRES: linear solver used to find an approximate solution to Ax=b from a Krylov subspace
- Many modeling and physics applications use discretizations of partial differential equations that require GMRES or some other Krylov solver.
- Belos: linear solvers package in Trilinos
- Kokkos and Kokkos Kernels: Portable parallel linear algebra software for GPUs
- My work:
  - Code a new adapter to use Kokkos as the linear algebra backend for Belos solvers
  - Implement mixed precision operations
  - Test performance improvements on a single node with GPU

Algorithm: GMRES with Iterative Refinement (GMRES-IR)

Why incorporate lower precision data storage in Krylov solvers?

- Krylov solvers are typically memory-bound (data movement is more expensive than FLOPs).
- Cheaper data movement and floating-point operations.
- Take advantage of new hardware for low-precision computations (e.g. GPU tensor cores).

Algorithm: Use mostly single (32-bit float) precision with occasional double (64-bit) precisio

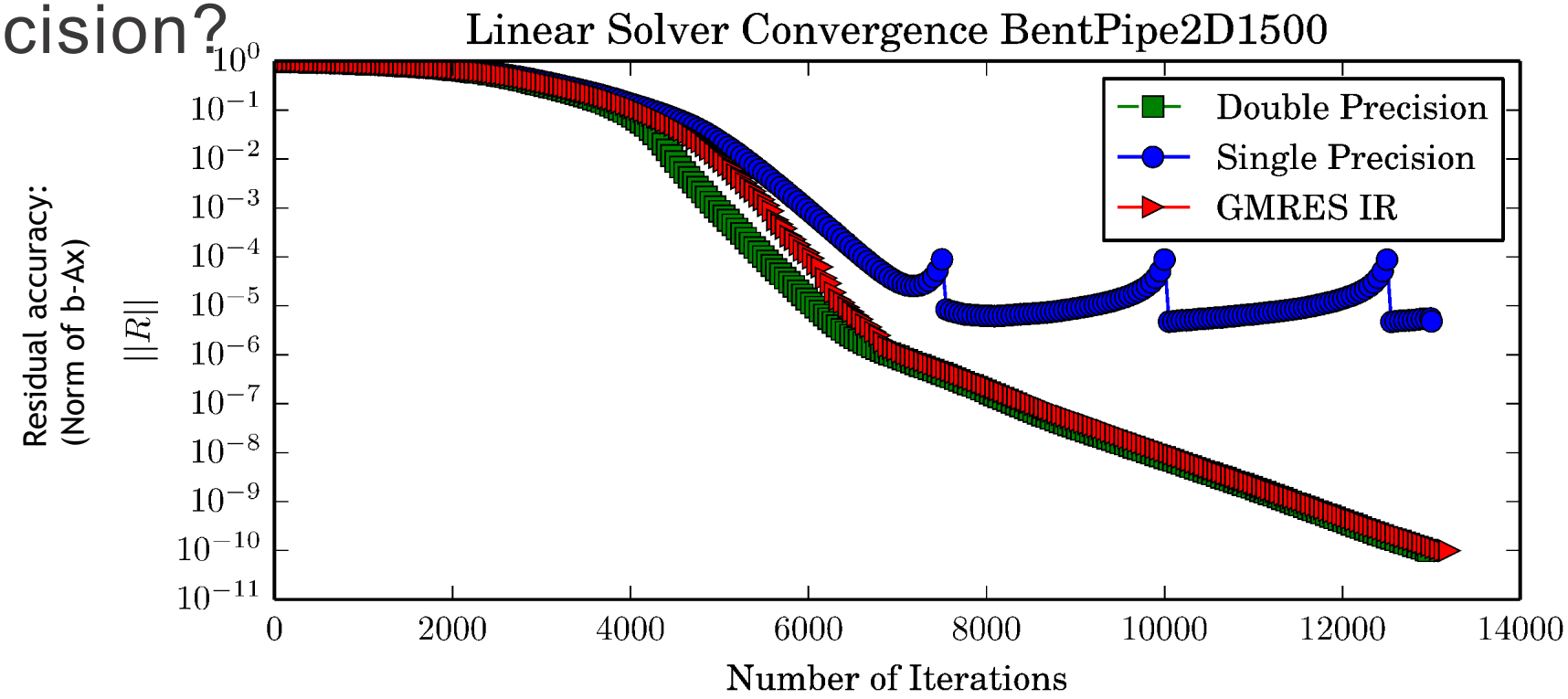
Algorithm 1 Iterative Refinement with GMRES Error Correction

1: $r_0 = b - Ax_0$ [double]
2: **for** $i = 1, 2, \ldots$ until convergence: **do**
3: Use GMRES(m) to solve $Au_i = r_i$ for correction u_i [single]
4: $x_{i+1} = x_i + u_i$ [double]
5: $r_{i+1} = b - Ax_{i+1}$ [double]
6: **end for**

Challenges:

- Lower precision computations result in more roundoff errors!
- Applications still need high level of accuracy in solutions

Is mixed-precision GMRES as accurate as double precision?



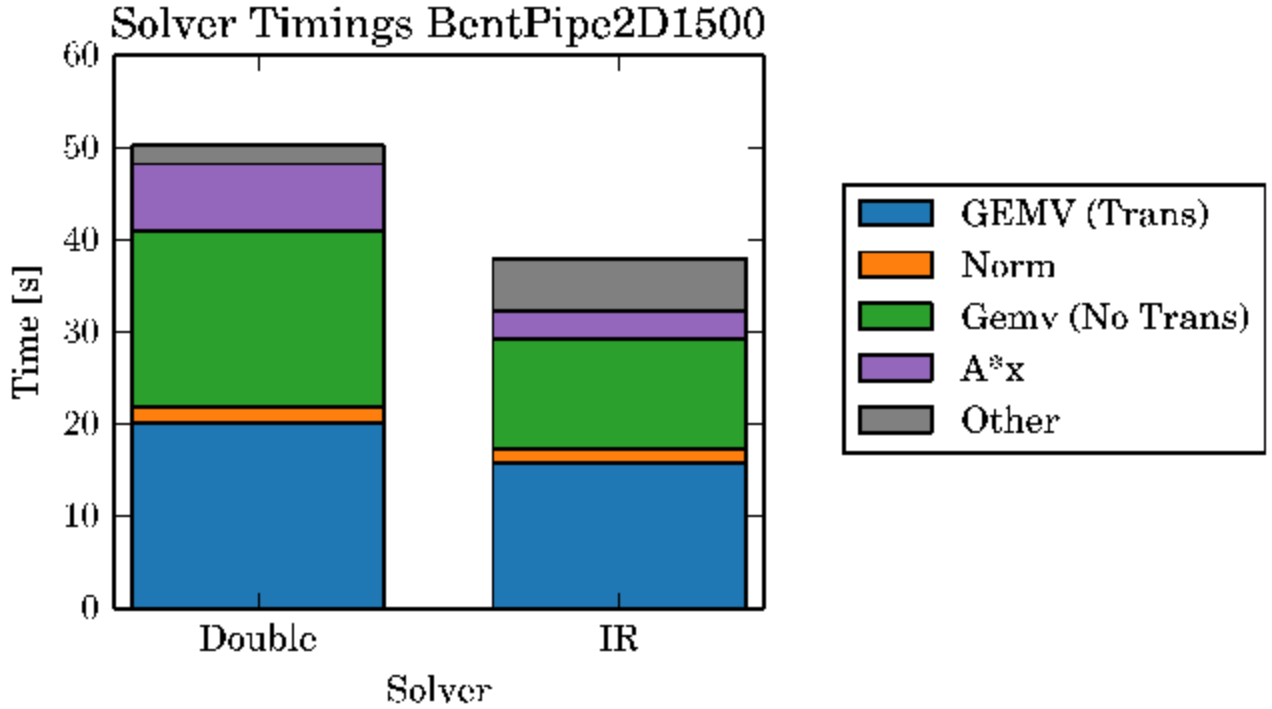
GMRES Double: 12967 iterations **GMRES-IR:** 13150 iterations

Matrix is 2D convection-diffusion problem over a 5-pt stencil. (Highly nonsymmetric.)
n = 2.25 million, number of non-zeros = 11,244,000

Running GMRES(50) to tolerance of 1e-10. (No preconditioning.)
For GMRES-IR: Residuals recomputed in double at each restart (each 50 iterations).
Tests run on a V100 GPU.

Is mixed-precision GMRES faster than double precision?

[Solver uses two steps of classical Gram-Schmidt (CGS(2)) orthogonalization]



	GMRES double	GMRES IR	Speedup
Total time:	**50.26**	**38.03**	**1.322**
Ortho: GEMV Trans	20.20	15.78	1.280
Ortho: GEMV No Trans	19.01	12.10	1.571
Ortho (norm)	1.71	1.49	1.152
A*x	7.33	2.95	2.484

(Timings do not include making an extra copy of the matrix A in single precision.)

Future Work:

- Test preconditioning (Block Jacobi, polynomial preconditioning, etc.…)
- Incorporate half (16-bit) precision
- Make available to Sandia applications using Trilinos and Tpetra

Thank you!

Related References:

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

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. This work was supported by the Department of Energy's Exascale Computing Project (ECP).