This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2022-4591C

xSDK Focus Effort Developing Multiprecision Numerics

EXASCALE COMPUTING PROJECT

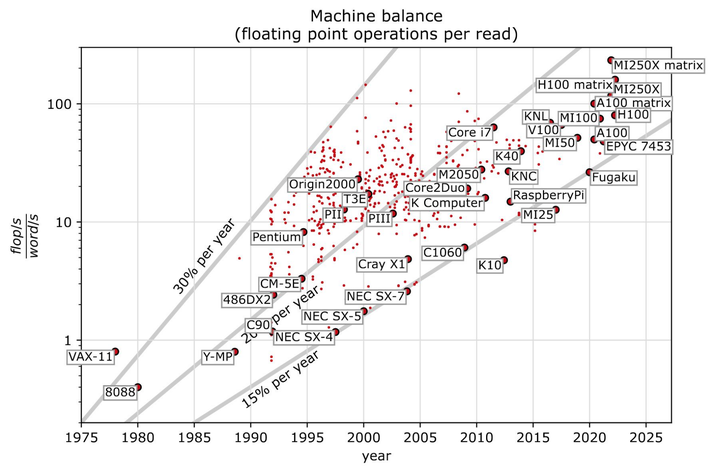
ECP Multiprecision Effort

Next-generation high performance math libraries need to exploit the compute power available in low precision cores

Recently, hardware manufacturers introduced low precision functionality like FP16 by integrating low precision cores, e.g., the NVIDIA Tensor cores. These, however, remain unused even for compute-intensive applications as long as high precision is employed for all arithmetic operations. While employing formats other than working precision can render attractive performance improvements, it also requires careful consideration of the numeric effects. On the other end of the spectrum, precision formats with higher accuracy than the hardware-supported FP64 can be effective in improving robustness and accuracy of numerical methods.

Challenging trend of the arithmetic performance growing faster than the processors' memory bandwidth

At the same time that hardware manufacturers introduce compute units providing orders of magnitude speedups in low precision computations, communication-intensive applications suffer from the memory bandwidth of architectures growing at a much lower pace than the arithmetic performance.

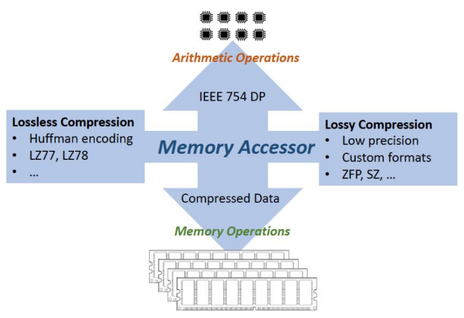


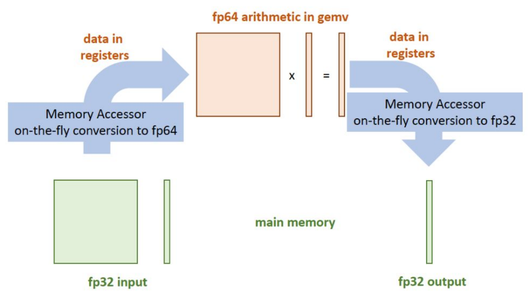
Machine Balance evolution, Mark Gates (UTK), updated 2022.

Mixed Precision BLAS

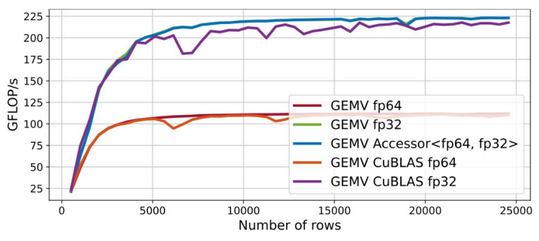
Memory Accessor

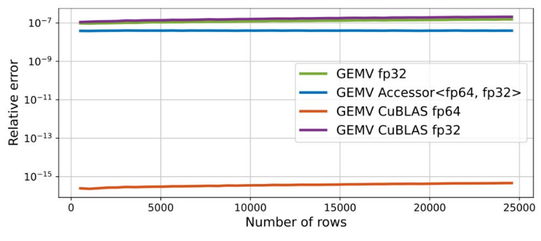
In response to the growing discrepancy between the arithmetic performance and the memory bandwidth, Ginkgo follows a concept that decouples the memory precision from the arithmetic precision and compresses all data in processor registers before issuing memory operations. This compression can be complex (compression formats) or simple (converting to lower precision). Incurring negligible overhead to the computational cost, the memory accessor enables for memory bound algorithms to maintain the performance of the low precision memory format why increasing the algorithm accuracy by using high precision in all arithmetic operations and in-register redictions.

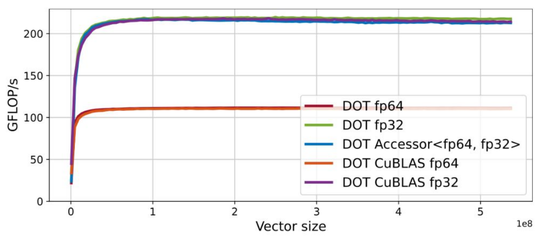


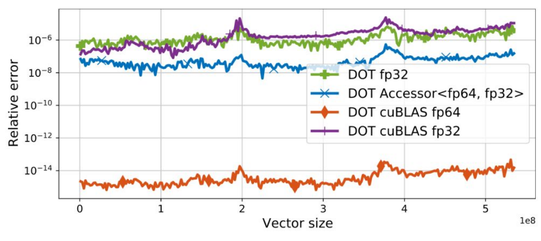


Mixed Precision BLAS using the Memory Accessor









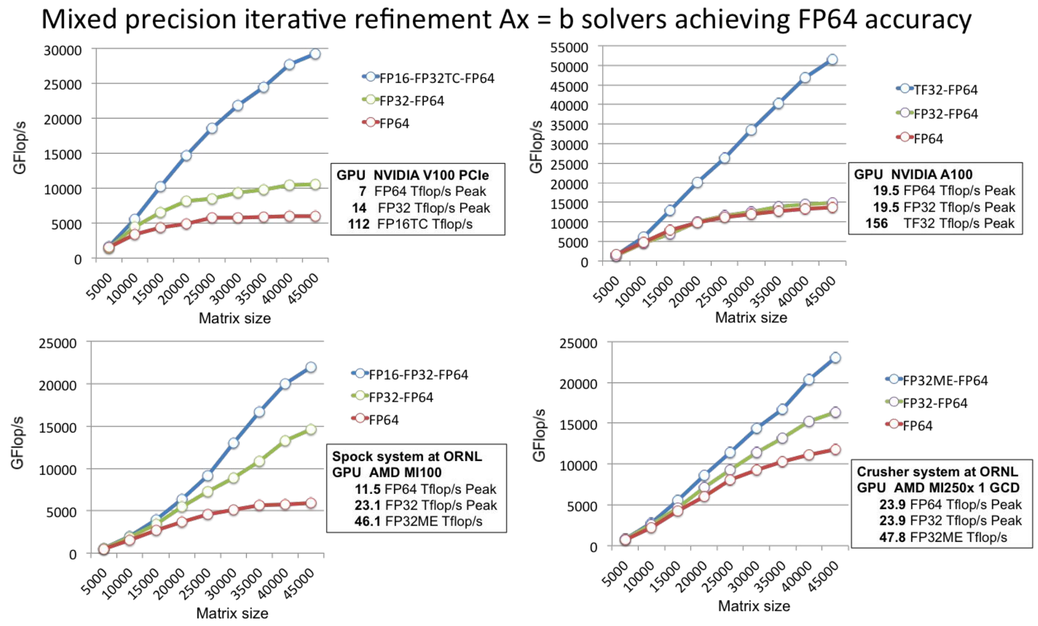
Mixed precision BLAS1/BLAS2 using the memory accessor matches the performance of single precision BLAS on the NVIDIA V100 GPU. Performing the arithmetic operations and local reductions in double precision, the routines have about one order of magnitude higher accuracy.

T Grützmacher, H Anzt, ES Quintana-Ortí: Using Ginkgo's memory accessor for improving the accuracy of memory-bound low precision BLAS, Software: Practice and Experience (2021).

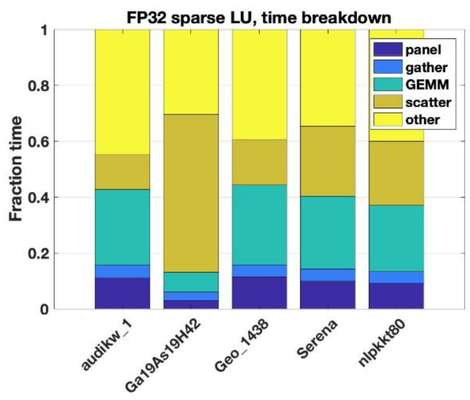
Mixed Precision Iterative Refinement

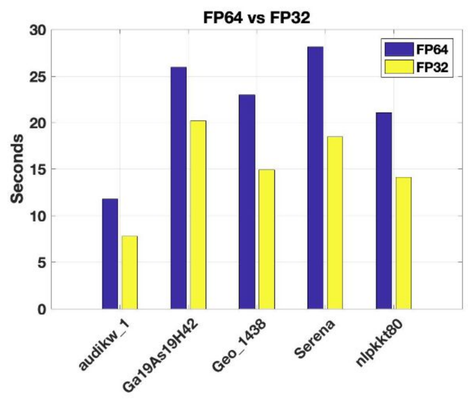
Mixed Precision Iterative Refinement Using Dense LU

Ax = b solvers can be accelerated significantly using lower precision (FP16/32 TC/ME) and the mixed-precision iterative refinement technique to retain FP64 accuracy. Illustrated are multiple times acceleration for dense matrices using MAGMA 2.6.2 on latest GPUs from NVIDIA and AMD.



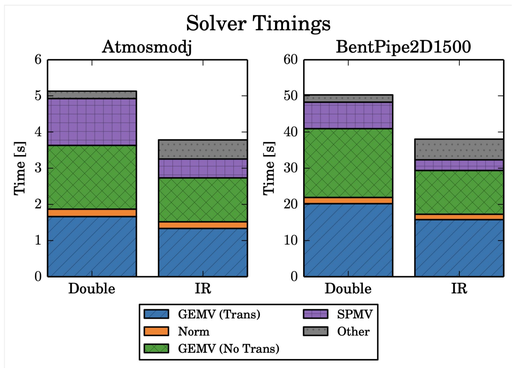
Mixed Precision Iterative Refinement Using Sparse LU





Mixed precision Iterative Refinement using GMRES

The Trilinos project recently deployed a mixed precision GMRES solver based on the idea of using a low precision GMRES inside a mixed precision iterative refinement solver. The approach has shown to provide performance benefits for relevant applications.



Solve times for GMRES(50) double (left) and IR (right) for the matrices Atmosmodj and BentPipe2D1500. Each bar represents total solve time, split up to give abreakdown of time spent in different kernels. The "Other" portion represents timing for small dense (non-GPU) operations and, for GMRES-IR, computingresiduals in fp64.

Multiprecision Landscape Assessment

In a collaborative effort of all project partners, we assembled a survey on numerical methods utilizing mixed precision arithmetic. A consolidated version of the report has been accepted for publication in the International Journal on High Performance Computing Applications.

arXiv.org > cs > arXiv:2007.06674

Computer Science > Mathematical Software

[Submitted on 13 Jul 2020]

A Survey of Numerical Methods Utilizing Mixed Precision Arithmetic

Ahmad Abdelfattah, Hartwig Anzt, Erik G. Boman, Erin Carson, Terry Cojean, Jack Dongarra, Mark Gates, Thomas Grützmacher, Nicholas J. Higham, Sherry Li, Neil Lindquist, Yang Liu, Jennifer Loe, Piotr Luszczek, Pratik Nayak, Sri Pranesh, Siva Rajamanickam, Tobias Ribizel, Barry Smith, Kasia Swirydowicz, Stephen Thomas, Stanimire Tomov, Yaohung M. Tsai, Ichitaro Yamazaki, Urike Meier Yang

ECP 5th Annual Meeting, April 2021

This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the 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.