Annotated Bibliography

Neil Lindquist

June 1, 2017

References

[1] Hartwig Anzt, Björn Rocker, and Vincent Heuveline. Energy efficiency of mixed precision iterative refinement methods using hybrid hardware platforms. *Computer Science - Research and Development*, 25(3):141–148, 2010.

> Experimental times for using mixed precision solvers on high end computers. The algorithm used utilized single precision to solve $Ac = r$ where c is the amount to increment x.

[2] Marc Baboulin, Alfredo Buttari, Jack Dongarra, Jakub Kurzak, Julie Langou, Julien Langou, Piotr Luszczek, and Stanimire Tomov. Accelerating scientific computations with mixed precision algorithms. *CoRR*, abs/0808.2794, 2008.

> The authors studied performance of iterative linear algebra solvers by doing the first few refinement passes with single precision floats, before using double precision to get the final awnser. They found that doing the first passes with single precision floats was significantly faster than using only double precision.

[3] Alfredo Buttari, Jack Dongarra, Jakub Kurzak, Piotr Luszczek, and Stanimir Tomov. Using mixed precision for sparse matrix computations to enhance the performance while achieving 64-bit accuracy. *ACM Trans. Math. Softw.*, 34(4):17:1–17:22, July 2008.

> The authors tried using both single and double precision in an iterative refinement algorithms to try to increase speed while retaining accuracy. The algorithms were modified to call a single precision linear solver to solve $Ac = r$ where c in amount to increment x. PCG and GMRES were both modified in this way.

[4] Alfredo Buttari, Jack Dongarra, Julie Langou, Julien Langou, Piotr Luszczek, and Jakub Kurzak. Exploiting the performance of 32 bit floating point arithmetic in obtaining 64 bit accuracy (revisiting iterative refinement

for linear systems). In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, SC '06, New York, NY, USA, 2006. ACM.

Single precision floats are able to be processed faster and the communication is much cheaper that working with double precision floats. The authors mixed single and double precision in each refinment step. They found that if factorization, forward substitution and backward substitution are do in single precision while the residual and update to the solution are done in double precision then the iterative refinement will produce the same accuracy than if double precision was used exclusively (as long as the matrix is not too badly conditioned).

[5] Alfredo Buttari, Jack Dongarra, Julie Langou, Julien Langou, Piotr Luszczek, and Jakub Kurzak. Mixed precision iterative refinement techniques for the solution of dense linear systems. *Int. J. High Perform. Comput. Appl.*, 21(4):457–466, November 2007.

The authors compared using strictly double precision to using mixed single and double precision in iterative linear equation solvers with dense matrices. Except for small problems, the mixed precision solver was able to solve faster. The algorithm in question first solved the system in single precision then refined it to double precision quality.

[6] Wei-Fan Chiang, Mark Baranowski, Ian Briggs, Alexey Solovyev, Ganesh Gopalakrishnan, and Zvonimir Rakamarić. Rigorous floating-point mixed-precision tuning. *SIGPLAN Not.*, 52(1):300–315, January 2017.

The authors address general improvements for mixing precision of floats in algorithms (not just iterative refinement solvers).

[7] James Demmel, Yozo Hida, William Kahan, Xiaoye S. Li, Sonil Mukherjee, and E. Jason Riedy. Error bounds from extra-precise iterative refinement. *ACM Trans. Math. Softw.*, 32(2):325–351, June 2006.

The authors address calculating the error bounds when the precision used to calculate the residual is higher than the precision used for the rest of the calculations in an iterative refinement algorithm.

[8] Samuel A. Figueroa. When is double rounding innocuous? *SIGNUM Newsl.*, 30(3):21–26, July 1995.

The author proves that doing an arithmetic operation on single precision floats in double precision then converting it to single precision is the same as doing the operation in single precision.

[9] Dominik Göddeke, Robert Strzodka, and Stefan Turek. Performance and accuracy of hardware-oriented native-, emulated-and mixed-precision solvers in fem simulations. *Int. J. Parallel Emerg. Distrib. Syst.*, 22(4):221–256, January 2007.

The authors address using mixed precision in iterative refinement algorithms, as well as emulating higher precision with a pair of single precision floats. The authors also discuss the ability of different precisions of floating point operations on various hardware.

[10] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.

The author takes a detailed look at calculating error bounds for floating point arithmetic and algorithms. Additionally, there is some discussion on how to minimize error in said algorithms.

[11] J. D. Hogg and J. A. Scott. A fast and robust mixed-precision solver for the solution of sparse symmetric linear systems. *ACM Trans. Math. Softw.*, 37(2):17:1–17:24, April 2010.

The author anaylizes uses a mixed precision strategy utilizing fallbacks to more accurate algorithms to reach the user's requested precision. It starts with a single precision solve of $Ax = b$, then applies mixed-precision iterative refinement, then falls back to mixed-precision FGMRES, if target precision is not yet met then it uses a double precision solver on $Ax = b$ followed by double precision iterative refinement, then uses double-precision FGMRES. If target precision is not met a this point, then algorithm results in an error.

[12] Claude-Pierre Jeannerod and Siegfried M. Rump. Improved error bounds for inner products in floating-point arithmetic. *SIAM Journal on Matrix Analysis and Applications*, 34(2):338–344, 2013.

The authors improve on the error bound for inner products from Higham's *Accuracy and Stability of Numerical Algorithms*.

[13] John Jenkins, Eric R. Schendel, Sriram Lakshminarasimhan, David A. Boyuka, II, Terry Rogers, Stephane Ethier, Robert Ross, Scott Klasky, and Nagiza F. Samatova. Byte-precision level of detail processing for variable precision analytics. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 48:1–48:11, Los Alamitos, CA, USA, 2012. IEEE Computer Society Press.

The authors are attempting to reduce IO bottlenecks by not sending all bytes of a double-precision float when doing calculations, resulting in a lower precision value.

[14] Toyohisa Kaneko and Bede Liu. On local roundoff errors in floating-point arithmetic. *J. ACM*, 20(3):391–398, July 1973.

The authors find the bound on error in float addition with a single precision accumulator with guard bits. They found that accuracy is almost as high as if a double precision accumulator had been used.

[15] Xiaoye S. Li, James W. Demmel, David H. Bailey, Greg Henry, Yozo Hida, Jimmy Iskandar, William Kahan, Suh Y. Kang, Anil Kapur, Michael C. Martin, Brandon J. Thompson, Teresa Tung, and Daniel J. Yoo. Design, implementation and testing of extended and mixed precision blas. *ACM Trans. Math. Softw.*, 28(2):152–205, June 2002.

The authors address the design decisions of using higher precision internally in BLAS, as well as some related features. A quote of note: "...Intel processors or their AMD and Cyrix clones, are designed to run fastest performingarithmetic to the full width, 80-bits, of their internal registers. These computers confer some benefits of wider arithmetic at little or no performancepenalty."

[16] Tran-Thong and Bede Liu. Floating point fast fourier transform computation using double precision floating point accumulators. *ACM Trans. Math. Softw.*, 3(1):54–59, March 1977.

The author looks at using a double precision accumulator when working with single precision data in a Fast Fourier Transformation.

[17] Eiji Tsuchida and Yoong-Kee Choe. Iterative diagonalization of symmetric matrices in mixed precision and its application to electronic structure calculations. *Computer Physics Communications*, 183(4):980–985, apr 2012.

The author looks at using mixed precision when iteratively diagonalizing a symmetric matrix.