



# Investigation into the mixed precision linear solver HSL\_MA79

HS Thorne

October 2016

©2016 Science and Technology Facilities Council



This work is licensed under a [Creative Commons Attribution 3.0 Unported License](https://creativecommons.org/licenses/by/3.0/).

Enquiries concerning this report should be addressed to:

RAL Library  
STFC Rutherford Appleton Laboratory  
Harwell Oxford  
Didcot  
OX11 0QX

Tel: +44(0)1235 445384  
Fax: +44(0)1235 446403  
email: [libraryral@stfc.ac.uk](mailto:libraryral@stfc.ac.uk)

Science and Technology Facilities Council reports are available online at: <http://epubs.stfc.ac.uk>

ISSN 1358-6254

Neither the Council nor the Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigations.

# Investigation into the mixed precision linear solver HSL\_MA79

H. Sue Thorne<sup>12</sup>

## ABSTRACT

In software development, a large amount of work is done to ensure that the resulting code is *efficient*, where the wall clock execution time is the normal measure for efficiency. In the future, there will be an increasing desire for codes to be *energy efficient*, that is, the amount of energy consumed whilst the code is run is minimised.

The use of mixed precision is one proposed methodology for reducing both wall clock execution times and energy consumption of some codes. In this report, we consider the HSL code HSL\_MA79, which is a mixed precision code for solving linear systems of the form  $Ax = b$ , where  $x$  is the unknown. We compare execution times and energy consumption of the code for a number of different matrices  $A$  in a bid to reveal whether there are classes of sparse, symmetric matrix problems where the mixed precision approach may be advantageous. In particular, we will consider the application area that the matrix is from.

This work forms part of the EPSRC Service Level Agreement funded Software Outlook project.

**Keywords:** Energy Efficient Computing, Intel Xeon (IvyBridge), Linear Solver, Single Precision, Double Precision, Mixed Precision, CCP

---

September 8, 2016

---

---

<sup>1</sup>Scientific Computing Department, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Campus, Didcot, OX11 0QX, UK.

<sup>2</sup>Correspondence to: sue.thorne@stfc.ac.uk

# 1 Background and motivation

The desire for scientific codes to be *efficient* is governed both by wanting to minimise the execution time and, increasingly, trying to reduce the amount of energy consumed whilst running the code. With the move from petascale computing to exascale computing on the horizon and the associated requirement to increase the floating-point operation rate by a factor of 100 whilst only increasing the power usage by a factor of 10, *energy efficient computing* is raising its head in importance. Additionally, some HPC service providers are proposing to base their service fees on the amount of energy consumed by the user rather than a metric purely based on the time and number of cores used.

In this report, we consider the use of a mixed precision approach to solving linear systems of the form  $Ax = b$ , where  $A$  is a real, sparse, symmetric matrix with  $n$  rows and columns, and  $nz$  non-zero entries ( $nz \ll n^2$ ),  $b$  is a real vector of length  $n$  and  $x$  is the vector we wish to compute. Such matrix problems arise in a large number of applications and it is not unusual to need to solve a sequence of similar yet slightly different linear systems. Our investigations will centre around the HSL [10] code HSL\_MA79 [9], which is a mixed precision solver for problems of the form  $AX = B$ , where  $A$  is as above but  $X$  and  $B$  can have more than one column.

In Section 2, we outline the mixed precision solver HSL\_MA79 and we outline our benchmark tests in Section 3. The results of our tests can be found in Section 4, with our conclusions in Section 5.

## 2 Mixed precision solver HSL\_MA79

When solving  $Ax = b$ , HSL\_MA79's default method for measuring the accuracy of the computed  $x$  is the scaled norm

$$\beta = \frac{\|Ax - b\|_\infty}{\|A\|_\infty \|x\|_\infty + \|b\|_\infty}. \quad (1)$$

The smaller the value of  $\beta$  for fixed  $A$  and  $b$ , the more accurate the computed value of  $x$  is with respect to this scaled norm.

An outline of the method used within HSL\_MA79 is given in Algorithm 1. The user sets the precision of the initial factorisation, `prec`, the choice of linear solver, `solver`, and `accuracy`, the value that  $\beta$  should not exceed for the method to successfully terminate.

HSL\_MA79 incorporates two different linear solvers for the factorisation of sparse, symmetric matrices: MA57 and HSL\_MA77. MA57 is a serial, in-core code whilst HSL\_MA77 is a multi-threaded and out-of-core code. Given the nature of our test matrices (Section 3) and the fact that we are concerned with the effect of using mixed precision over that of using just single or double precision, we chose to only use MA57.

Iterative refinement is a cheap iterative method that takes an approximation to the solution  $x$  and aims to incrementally improve the solution. Whilst each iteration is cheap, it may be slow to converge. If iterative refinement is not converging very quickly, HSL\_MA79 switches to using the iterative method FGMRES, which, on average, is a lot more expensive than iterative refinement with respect to the cost of each iteration but it normally has better convergence properties. Further details about the methods used within HSL\_MA79 can be found in [9]. By default, HSL\_MA79 will switch to using a double precision factorisation if the required level of accuracy is not reached: in these tests, we have turned off this default setting by setting `fallback_double = .FALSE..` All other default controls were used.

## 3 Benchmark tests

In Table 1, we list the matrices considered in our tests. The matrices were all obtained from the University of Florida Sparse Matrix Collection [5] and have been grouped according to their classification within [5]. For our test set, we have chosen real, symmetric problems where the number of rows/columns,  $n$ , are fairly modest and the matrices are all sparse, that is,  $nz \ll n^2$ . We always set the right-hand side,  $b$ , to be a vector of length  $n$  with entries all equal to 1. For each matrix, we compare four different methods:

---

**Algorithm 1** HSL\_MA79 outline

---

```
Analyse matrix  $A$  using MA57
Choose solver and initialise precision prec
loop
  Factor  $A$  using solver with prec precision
  Use computed factors to (approximately solve  $Ax = b$ )
  Compute scaled residual  $\beta$ 
  if  $\beta \leq \text{accuracy}$  then
    return  $x$ 
  end if
  Apply iterative refinement and recompute  $\beta$ 
  if  $\beta \leq \text{accuracy}$  then
    return  $x$ 
  end if
  Apply FGMRES and recompute  $\beta$ 
  if prec = SINGLE and fallback_double then
    prec = DOUBLE
  else
    Return  $x$  with warning
  end if
end loop
```

---

- S Factor  $A$  using single precision and use **accuracy** =  $2 \times 10^5$  (i.e., no iterative refinement or FGMRES);
- D Factor  $A$  using double precision and use **accuracy** =  $2 \times 10^5$  (i.e., no iterative refinement or FGMRES);
- M1 Factor  $A$  using single precision and use **accuracy** =  $5 \times 10^{-15}$ ;
- M2 Factor  $A$  using single precision and use **accuracy** =  $5 \times 10^{-12}$ .

For each matrix and method, we report

$t_f$ , the time to factor  $A$  (only reported for Methods S and D);

$t_{afs}$ , the time to analyse  $A$ , factor  $A$  and compute  $x$  to desired level of accuracy;

$E_{tot}$ , the total energy consumption to read in  $A$ , analyse and factor  $A$ , and compute  $x$ ;

$\beta$ , the scaled norm (1) for the computed  $x$ ;

$ItRef$ , the number of iterative refinement iterations performed;

$ItFG$ , the number of FGMRES iterations performed.

All of the tests were run on Neale, a system provided by STFC's Hartree Centre [1]. Neale contains Intel Xeon (IvyBridge ES-2650v2) based nodes: each node has 8 cores (16 threads), the clock speed is 2.6GHz (3.6 GHz Max) and the L1 and L2 Cache sizes are 32 kB and 256 kB, respectively. Each node has 64 GB of memory. Energy monitoring on Neale uses the PAPI (Performance API) interface to begin an energy monitoring thread on an individual node while the test runs. Results consist of a trace of total power consumption from the A2 node and memory domains and the energy consumption is approximated from this power trace.

In all of the figures, the test problems have been ordered in the same order as in Table 1 and vertical lines have been used to separate different classes of problems.

Class	Name	$n$	$nz$
Structural	apache1	80800	542184
	apache2	715176	4817870
	bcsstk36	23052	1143140
	bmw3.2	227362	11288630
	bmw7st.1	141347	7339667
	crankseg.1	52804	10614210
	ct20stif	52329	2698463
	dawson5	51537	1010777
	F2	71505	5294285
	hood	220542	9895422
	msc23052	23052	1154814
	nasasrb	54870	2677324
	oilpan	73752	3597188
	pwtck	217918	11634424
	s3dkq4m2	90449	4820891
	s4dkt3m2	90449	3753461
	ship_003	121728	8086034
	shipsec8	114919	6653399
	thread	29736	4470048
	vanbody	47072	2329056
Network	case39	40216	1042160
	TSOPF_FS_b39_c30	120216	3121160
Model order reduction	boneS01	127224	6715152
	filter3D	106437	2707179
	rail.79841	79841	553921
	t3dl	20360	509866
Comp. fluid dynamics	cf1	70656	1828364
	cf2	123440	3087898
	copter2	55476	759952
	parabolic.fem	525825	3674625
	shallow_water1	81920	327680
	stokes128	49666	558594
Optimisation	a0nsdsil	80016	355034

Class	Name	$n$	$nz$
Optimisation	a5esindl	60008	255004
	blockqp1	60012	640033
	boyd2	466316	1500397
	brainpc2	27607	179395
	c-56	35910	380900
	cont-201	80595	438795
	cont-300	180895	988195
	dixmaanl	60000	299998
	jnlbrng1	40000	19920
	k1_san	67759	559774
	ncvxbqp1	50000	349968
	torsion1	40000	197608
Circuit sim.	G2.circuit	150102	726674
Economics	finan512	74752	596992
Comp. graphics	Andrews	60000	760154
Acoustics	qa8fk	66127	1660579
Thermal	thermal1	82654	574458
	thermomech.dM	204316	1423116
Electromagn.	tmt_sym	726713	5080961
2D/3D	aug3dcqp	35543	128115
	cant	62451	4007383
	d_pretok	182730	1641672
	darcy003	389874	2101242
	Dubcova2	65025	1030225
	Dubcova3	146689	3636649
	ecology1	1000000	4996000
	ecology2	999999	4995991
	helm2d03	392257	2741935
	helm3d01	32226	428444
	tuma1	22967	87760
	turon_m	189924	1690876
	wathen120	36441	565761
Materials	crystk03	24696	1751178

Table 1: The test matrices considered in our tests along with the number of rows/columns,  $n$ , and the number of non-zero entries,  $nz$ . The class of each matrix is also given.

## 4 Results and discussion

In Table 2 (at the end of the report), we list the values of  $t_{afs}$ ,  $E_{tot}$ ,  $\beta$ ,  $ItRef$  and  $ItFG$  for each test problem and method. In Figure 1, we compare the value of  $t_{afs}$  for Methods S and D. As expected, using single precision is faster than double precision for most test problems. For those classed as structural problems, there is between a 13 and 30% reduction in  $t_f$  but, as seen in Figure 2, the value of  $\beta$  is roughly  $10^8$  times larger. With regards the optimisation problems, for 7 out of the 13 problems there is no difference in the value of  $t_{afs}$  but the value of  $\beta$  is about  $10^8$  times larger when single precision is used. It should also be noted that the values of  $\beta$  from Method D for 5 of the optimisation problems were also several orders of magnitude larger than that of most other problems in our large test set, which indicates the difficult nature of these problems.

Factorising  $A$  using double precision and computing  $x$  using these factors resulted in  $\beta$  being larger than  $5 \times 10^{-15}$  for 11 out of the 65 test problems. For all of these 11 test problems, Method M1 computed an  $x$  for which  $\beta$  was smaller than  $5 \times 10^{-15}$  and, hence, the mixed precision methodology produced a more accurate value of  $x$  for these problems. The most extreme case was the optimisation test problem k1\_san, Method D resulted in  $\beta = 6.5 \times 10^{-7}$  whilst Method M1 achieved  $\beta < 5 \times 10^{-15}$  with 1 iterative refinement iteration and 12

FGMRES iterations. Thus, even without time and energy consumption considerations, a mixed precision approach may be advantageous over a double precision factorisation. This also makes it an attractive approach for when higher than double precision accuracy is required but the factorisation can only be done using double precision arithmetic.

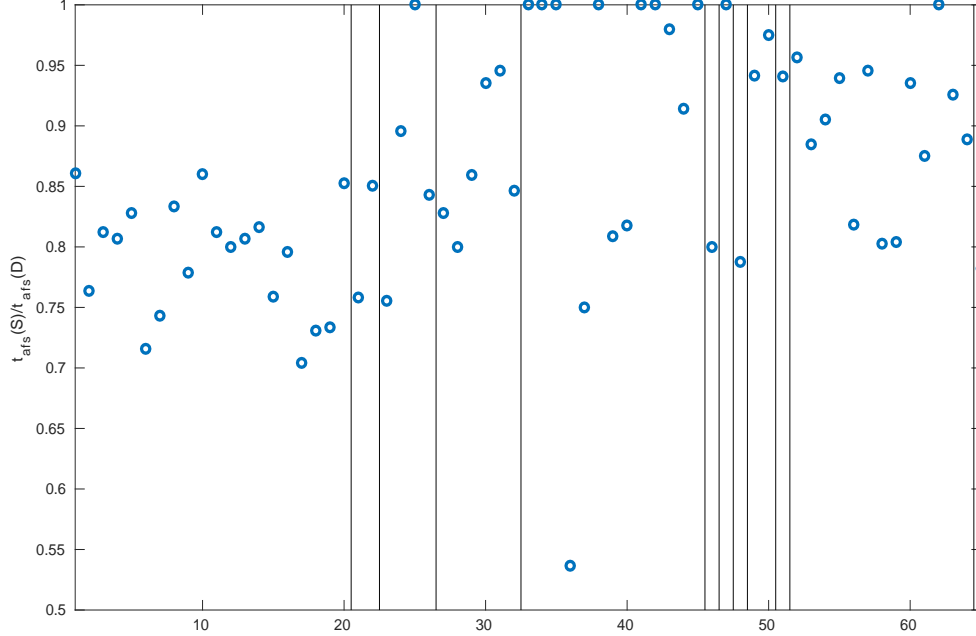


Figure 1: Value of  $t_{afs}(S)/t_{afs}(D)$ .

In Figures 3 and 4, we compare the execution times and energy consumption values for Methods D and M1. For the 11 problems where Method D failed to calculate an  $x$  satisfying  $\beta < 5 \times 10^{-15}$ , Method M1 took slightly less time for 3 of the problems, the same amount of time for 2 of the problems and between 15 and 45% longer for the remaining 6 problems. If it is important that  $\beta < 5 \times 10^{-15}$ , then this increase in time will generally be acceptable. The energy consumption ratios are slightly larger than those of the execution times but, again, the values are almost certainly within an acceptable level. Method M1 failed to compute an  $x$  satisfying  $\beta < 5 \times 10^{-15}$  for 9 of the problems. For 8 of these problems, the method had longer execution times than when Method D was used: this is due to the large number of FGMRES iterations. For our two network problems, Method M1 failed to reach the desired level of accuracy with the execution times and energy consumption values increased by at least 45% and, hence, if the method had been allowed to fallback to using double precision, then the mixed precision method would take roughly 2.5 times the time and energy to calculate an acceptable  $x$ . This highlights that, if very high accuracy is required, then there may be some classes of problems where it is best to stay with using double precision factorisations. For the computational fluid dynamics problems, there was always an advantage to using Method M1 over Method D with respect to execution time; there was little difference when energy consumption is considered. For the 12 optimisation problems, Method M1 failed to reach the desired accuracy for one problem but, for the 6 problems where Method D performed well, there was little difference in time and energy consumption for 4 of the problems. Certainly, when high accuracy is required for linear systems from optimisation, mixed precision should be seriously considered. For the 13 2D/3D problems, the execution times differ by up to 20% (increase and decrease) but Method M1 achieved the required accuracy for the 3 problems where Method D failed; there is normally a slight increase in energy consumption. In general, the results show

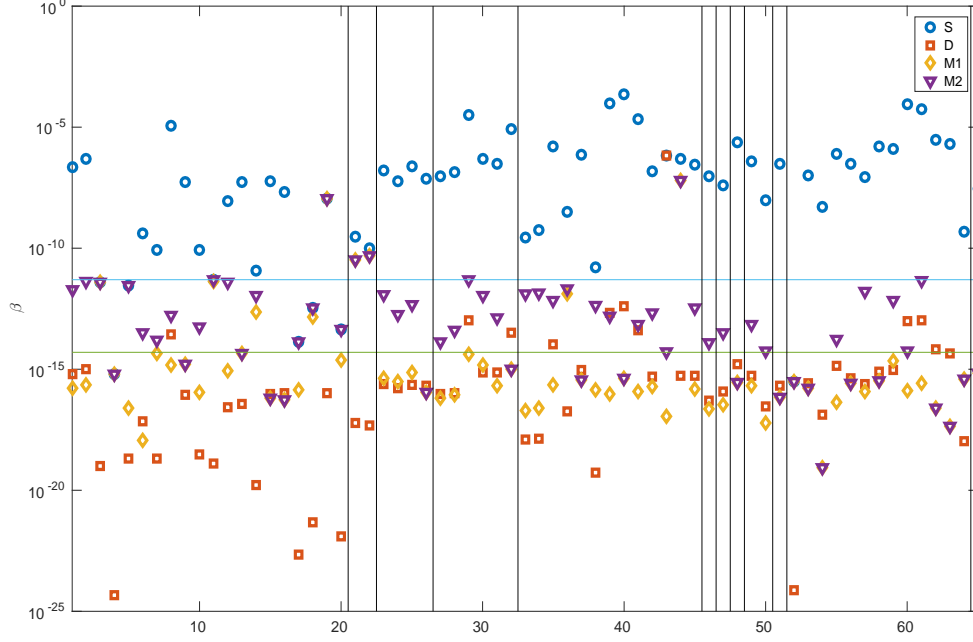


Figure 2: Value of  $\beta$  for Methods S, D, M1 and M2.

that, when high accuracy is required, the mixed precision approach may be of use and it is worth investigating whether it should be used for your particular application.

By the nature of some applications and the methods used to formulate the linear system, finding an  $x$  that satisfies  $\beta < 5 \times 10^{-15}$  is often overkill because the data involved has much lower levels of accuracy. If, instead, we require that  $\beta < 5 \times 10^{-12}$ , then we will expect the number of iterations to decrease and, hence, for execution times and energy consumption values to drop. In Figures 5 and 6, we observe that the execution times and energy consumption values are similar for the vast majority of problems but Method M2 only fails for four of the problems: again, there are failures for the two network problems. For the four structural problems where Method M1 fails but Method M2 is successful, the execution times for Method M2 are between 43 and 62% lower than Method M2 because the number of FGMRES iterations have dramatically dropped; the energy consumption values have dropped by between 23 and 46%.

In Figures 7 and 8, we compare Methods D and M2. Method D on failed to the reach the desired accuracy of  $\beta < 5 \times 10^{-12}$  for one problem (k1.san). Of the 65 test problems, Method M2 is successful for 61 of the problems and, of these problems, it has a negative impact on the execution times for just 16 of the problems (we consider there to be no negative impact for the problem that Method D failed on); the ratios for energy consumption are slightly larger. Hence, when lower levels of accuracy for are required for the solution of linear, mixed precision methods are even more attractive, particularly for structural, model order reduction and computational fluid dynamics problems.

## 5 Conclusions

We have seen that if a linear system is being solved by factoring the matrix using double precision accuracy, then switching to a mixed precision approach may be beneficial, particularly for execution times. For problems where the solution needs calculating to a very high accuracy but the double precision factorisation does not compute

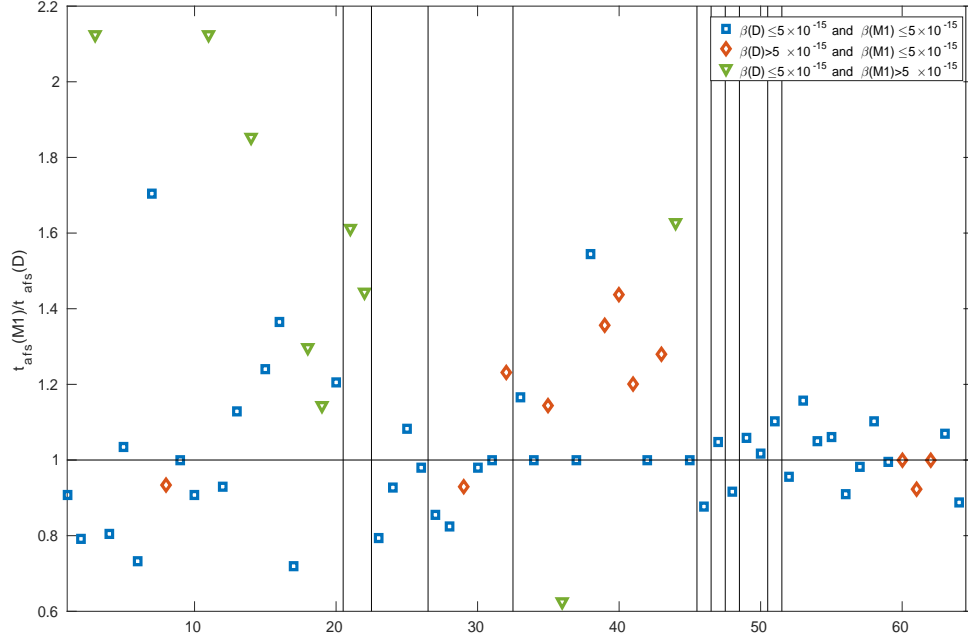


Figure 3: Value of  $t_{afs}(M1)/t_{afs}(D)$ .

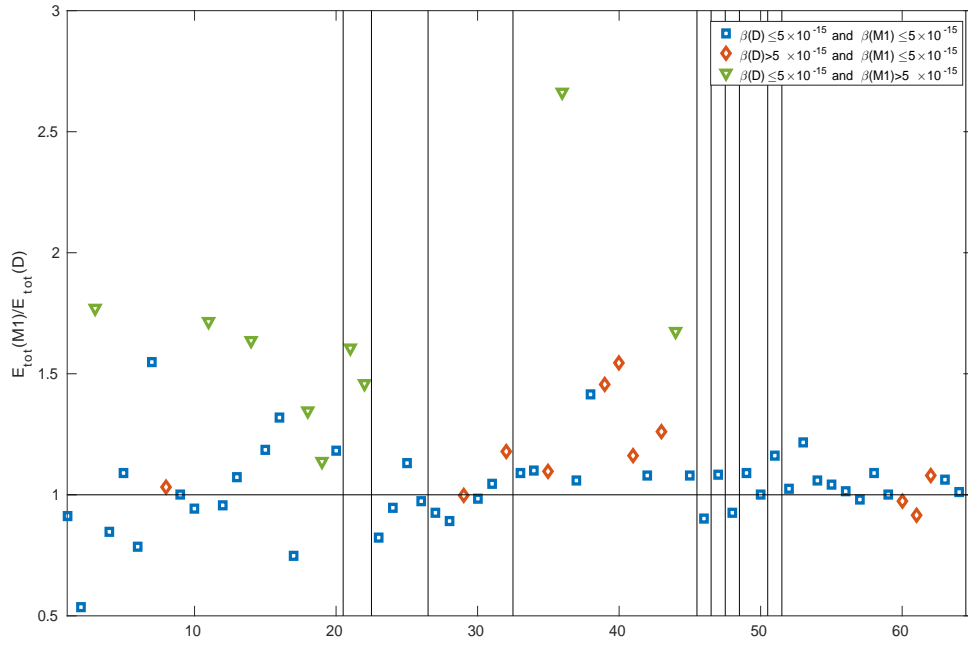


Figure 4: Value of  $E_{tot}(M1)/E_{tot}(D)$ .

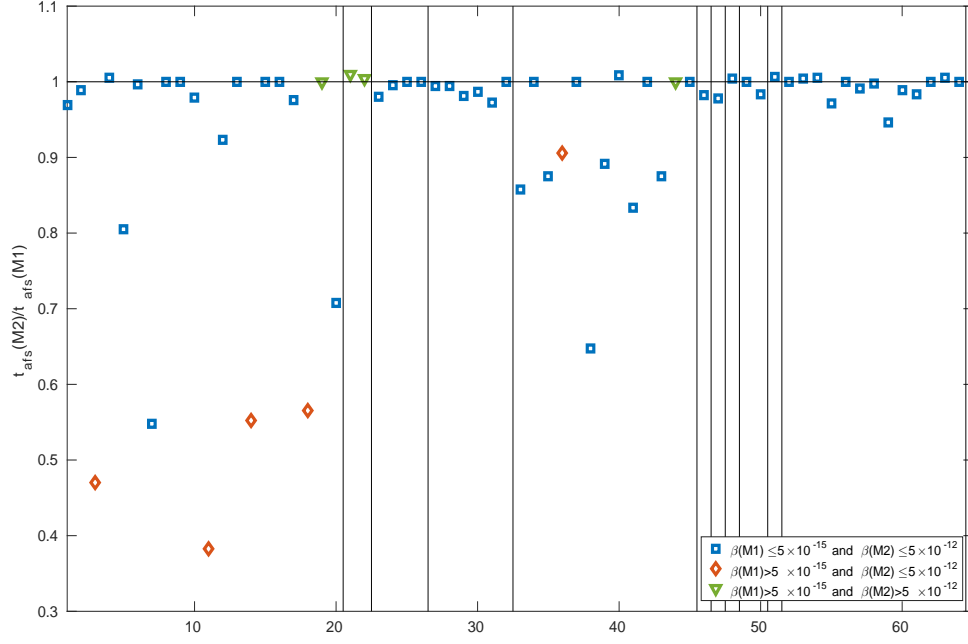


Figure 5: Value of  $t_{afs}(M2)/t_{afs}(M1)$ .

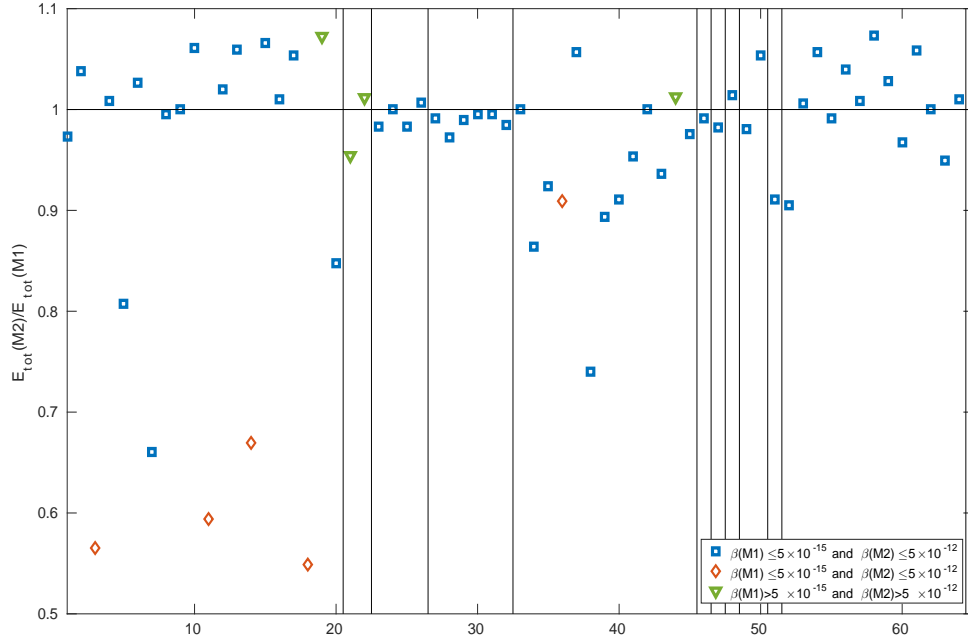


Figure 6: Value of  $E_{tot}(M2)/E_{tot}(M1)$ .

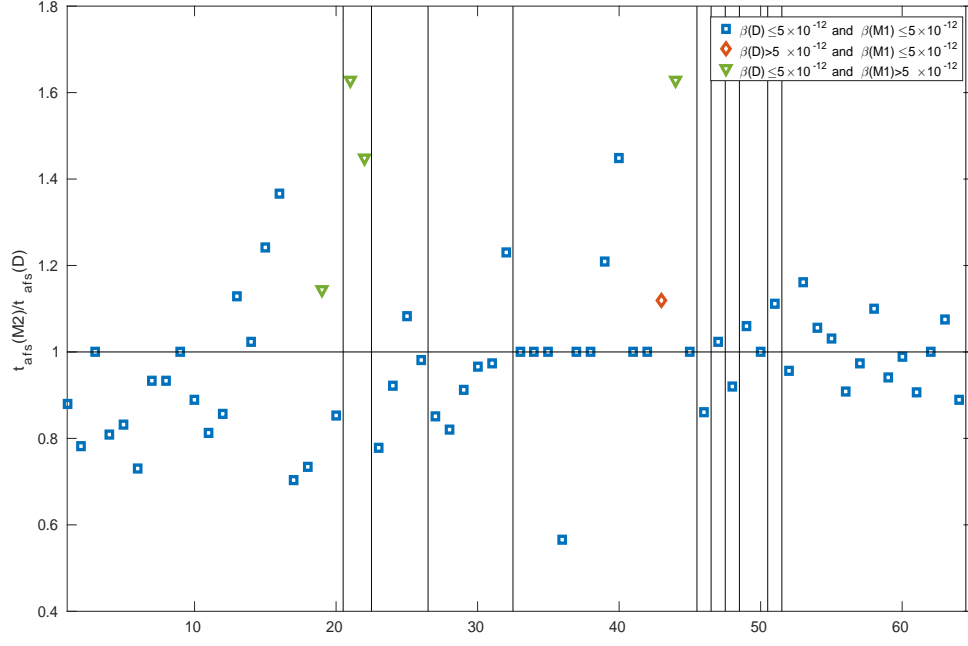


Figure 7: Value of  $t_{afs}(M2)/t_{afs}(D)$ .

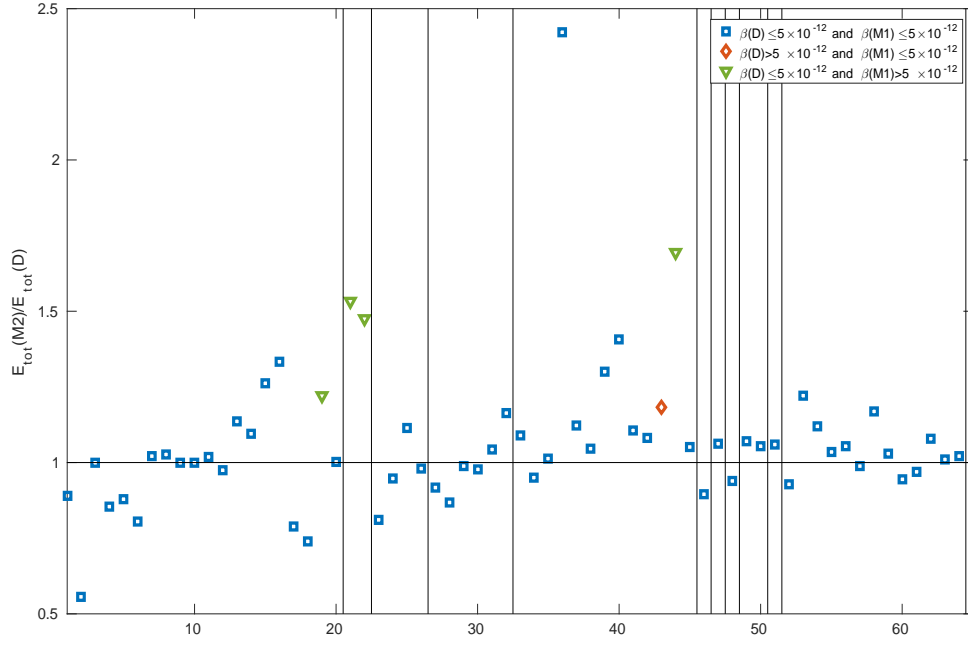


Figure 8: Value of  $E_{tot}(M2)/E_{tot}(D)$ .

a good enough solution, the mixed precision approach is very attractive. When a lower level of accuracy is required, the mixed precision approach becomes even more attractive for some problems. For some problems, a double precision factorisation may not be possible due to memory restrictions but a single precision factorisation is possible and the mixed precision approach could then be used. We should also note that iterative refinement did not converge quickly enough and FGMRES had to be used for many of the problems. FGMRES can be a very expensive method and cheaper Krylov-based methods may be available for certain classes of matrix, for example, the preconditioned conjugate gradient method for symmetric, sparse, positive definite matrices, which has well understood convergence properties [12]. For problems where a factorisation of  $A$  is not possible in either double or single precision due to memory or time constraints, there are a number of iterative method possibilities along with ways to improve the rate of convergence, for example, the use of preconditioners [6, 7, 12].

Below we list some of the available sparse linear solvers that contain methods for refining the solution  $x$ .

**HSL\_MA79** Either in-core and serial, or out-of-core and multi-threaded. Single and double precision versions available with iterative refinement and FGMRES

**MUMPS** Parallel code (MPI and OpenMP) for symmetric and unsymmetric problems. Single and double precision versions available with iterative refinement [2]

**PaStiX** Parallel code (MPI and OpenMP) for symmetric and unsymmetric problems. Single precision factorisation NOT available but has a number of iterative refinement methods available to improve accuracy of the solution [8, 3]

**SuperLU** Parallel code (MPI and OpenMP) for unsymmetric with single and double precision versions available with iterative refinement [11, 4].

The ideal linear solver would be an in-core parallel code that has both single and double precision versions and methods for refining  $x$  that include Krylov-subspace iterative methods. Such a code does not seem to be publicly available at the moment.

## References

- [1] See <http://www.hartree.stfc.ac.uk/hartree/>.
- [2] See <http://mumps.enseeiht.fr/index.php?page=home>.
- [3] See <http://pastix.gforge.inria.fr/>.
- [4] See [http://http://crd-legacy.lbl.gov/~xiaoye/SuperLU/#superlu\\_mt](http://http://crd-legacy.lbl.gov/~xiaoye/SuperLU/#superlu_mt).
- [5] T. DAVIS, *The University of Florida Sparse Matrix Collection*, 2007. <http://www.cise.ufl.edu/davis/techreports/matrices.pdf>.
- [6] H. C. ELMAN, D. J. SILVESTER, AND A. J. WATHEN, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford University Press, Oxford, 2005.
- [7] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, third ed., 1996.
- [8] P. HÉNON, P. RAMET, AND J. ROMAN, *PaStiX: A High-Performance Parallel Direct Solver for Sparse Symmetric Definite Systems*, *Parallel Computing*, 28 (2002), pp. 301–321.
- [9] 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 (2010), pp. 17:1–17:24.
- [10] HSL, *A collection of Fortran codes for large-scale scientific computation*. <http://www.hsl.rl.ac.uk/>, 2013.

- [11] X. S. LI, *An overview of SuperLU: Algorithms, implementation, and user interface*, ACM Trans. Math. Softw., 31 (2005), pp. 302–325.
- [12] Y. SAAD, *Iterative methods for sparse linear systems*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2003.

